Nucleosome positioning on chromatin using bioinformatics techniques

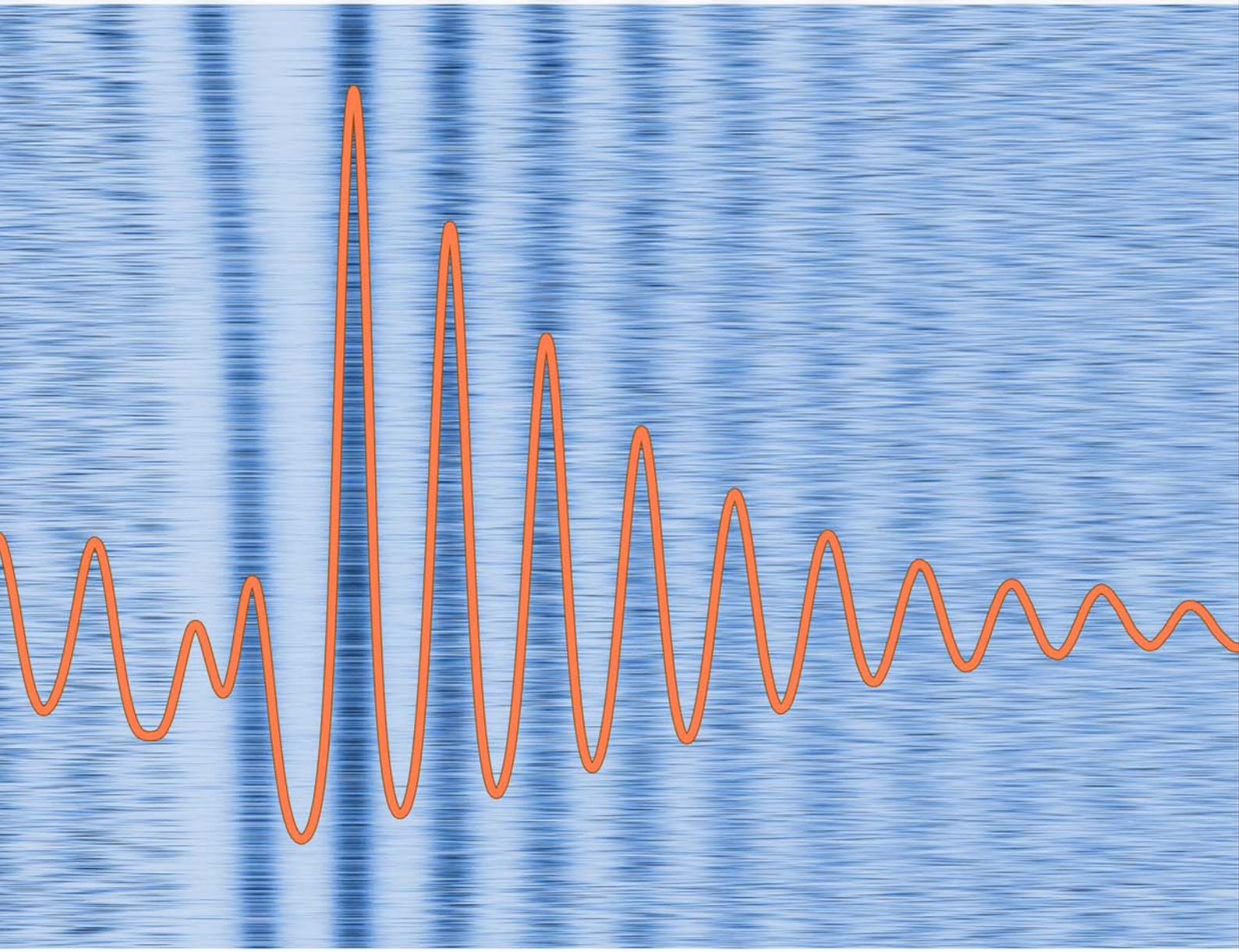
Posicionament de nucleosomes en cromatina emprant tècniques bioinformàtiques

Oscar Flores Guri

Oscar Flores Guri

PhD Thesis

Nucleosome positioning on chromatin using bioinformatics techniques



University of Barcelona

Faculty of Biology

Director: Modesto Orozco

PROGRAMA DE DOCTORAT EN BIOMEDICINA

Tesis realitzada en el grup de Modelatge Molecular i Bioinformàtica (MMB)

Programa conjunt de Biologia Computacional entre
el Institut de Recerca Biomèdica de Barcelona (IRB Barcelona)
i el Centre de Supercomputació de Barcelona (BSC-CNS)

POSICIONAMENT DE NUCLEOSOMES EN CROMATINA EMPRANT TÈCNIQUES BIOINFORMÀTIQUES

Memòria presentada per Oscar Flores Guri
per optar al grau de doctor per la Universitat de Barcelona

Director: Modesto Orozco López

Doctorand: Oscar Flores Guri

To my parents, Antonio and Consol

Als meus pares, Antonio i Consol

AGRAÏMENTS

Una tesi doctoral comença com un repte, continua com un sacrifici i acaba amb una gran il·lusió. Per això voldria agrair a la gent que m'ha motivat en el repte, que m'ha acompanyat en el sacrifici i que ara comparteixen aquesta il·lusió amb mi.

Dins els primers incloc aquelles persones del CMU Penyafort-Montserrat que em van transmetre la seva passió pel coneixement: en **Rafael**, la **Montse**, en **Kike**, n'**Àlex**, en **Jordi**, en **Tomeu**, i en especial la **Cèlia**, qui indirectament va ser la causant de que jo acabés en un doctorat del camp de la biologia.

La llista dels que m'han acompanyat durant aquests anys és molt més llarga:

Vull agrair a l'**Adam**, en **Nacho**, en **Jose**, en **Pedro**, l'**Antonella** les bones estones de desconnexió durant els dinars i la seva disposició d'ajudar i aguantar les meves idees de bomber. A la **Federica**, la **Michela**, l'**Ivan**, la **Nadine**, en **Pablo**, l'**Agustí** i més recentment l'**Antonija**, en **Hansel** i l'**Alexandra** les rialles al laboratori i les cerveses durant les *cool-offs*. Als membres més formals com l'**Alberto**, en **Pedro**, l'**Annalisa**, en **Guillem**, l'**Andreu** i tota l'altra gent que ha passat pel MMB els hi agraeixo les converses que hem mantingut, tan professionals com personals, al llarg d'aquests cinc anys. Vull agrair també a la gent del BSC, la **Sílvia**, en **Carles**, en **Ramón**, en **Pau** i a la resta de gent del *Life Science* els seus desplaçaments pels *group meetings* i les activitats "extraoficials" dels *retreats* de Núria. Agraeixo també a les noies del EBL, la **Chiara**, la **Montse**, la **Isabelle** la seva feina i la frescor del seu punt de vista experimental. A la resta de doctorands i altres membres del IRB, i molt en especial a l'**Albert**, l'**Abraham**, l'**Helena** i en **Giuseppe**, els hi vull agrair que hagin convertit el fet de treballar al IRB en una experiència que va molt més enllà del terreny professional o acadèmic. Vull donar les gràcies també als membres del meu TAC, en **David**, en **Josep Lluís** i en **Patrick** per la seva disposició i els seus consells tan a les reunions de seguiment com a fora.

Entre aquests companys de viatge, vull dedicar un agraïment especial a dues persones. La primera és en **Modesto**, per confiar en un informàtic que poc o res sabia de biologia quan va entrar i per la confiança i la llibertat que he sentit al llarg d'aquests anys. La segona és la **Özgen**, la meva companya infatigable de recerca amb qui he compartit penes i alegries, preocupacions professionals i personals, discussions i descobriments.

I finalment vull acabar aquest agraïment amb aquelles persones que mereixen estar en la meva tesi o en qualsevol altre moment de la meva vida. Persones que sempre hi són, encara que sigui a kilòmetres de distància, com la **Casandra** i en **Gonzalo**. Persones com la **Maria**, qui m'ha fet créixer tant aquests anys. Persones que són referència i escalfor a la vegada, com els meus **familiars**, els meus **germans**, la meva **tieta**, i en especial, els meus pares: l'**Antonio** i la **Consol**.

Per a ells dos, que han viscut la investigació amb tanta o més passió que jo, per aquesta i pels milers de raons més que no puc enumerar aquí, aquesta tesis doctoral està dedicada a ells.

TABLE OF CONTENTS

ABBREVIATIONS

bp	Base pairs
DAS	Distributed Annotation System
DNA	Deoxyribonucleic Acid
EBL	Experimental Bioinformatics Laboratory
FACS	Fluorescence-activated Cell Sorting
FFT	Fast Fourier Transform
HMM	Hidden Markov Models
LRs	Low coverage Regions
MD	Molecular Dynamics
MMB	Molecular Modeling and Bioinformatics
MNase	Micrococcal Nuclease
NFR	Nucleosome Free Region
NGS	Next Generation Sequencing
PCR	Polymerase Chain Reaction
PDB	Protein Data Bank
RNA	Ribonucleic Acid
TF	Transcription Factors
TFBS	Transcription Factor Binding Site
TSS	Transcription Start Site
TTS	Transcription Termination Sites

PREFACE

The present report illustrates my research work done in the period 2009-2014 in the Molecular Modeling and Bioinformatics Group at the Institute for Research in Biomedicine of Barcelona and the Barcelona Supercomputing Center. My main area of research during this time has been computational biology with a clear focus in the study of the chromatin structure and particularly nucleosome positioning using bioinformatic techniques.

Regarding this point, it is necessary to differentiate between the field of bioinformatics (*bio-*: life, *info-*: information, *-matics*: automatization) and programs which model a biological process in order to understand and predict it, more related to the field of computational biology (computational: using computers, bio-: life, -logos: reasoning). Bioinformatics would be more related to the technical manipulation of data meanwhile computational biology will refer to the use of computers to understand and predict biological problems.

The evolution of the research projects in the area has guided my research in three different topics:

i) Analysis of the DNA from a physical point of view, correlating chromatin structure with DNA function and its intrinsic properties,

ii) Development of tools to extract information on nucleosome position and dynamics from experimental data —mostly Next Generation Sequencing experiments— and

iii) In depth analysis of chromatin structure and dynamics in model systems (such as *S.Cerevisae, C.Elegans or, in a smaller extent, Human*)

Because of the broad nature of the topics covered in this dissertation I will give just an overview of the general concepts shared among the different studies and expand them in the results section if required. For purposes of clarity, the sections *Objectives*, *Publications* and *Discussion and conclusions* of this thesis will be divided in three chapters, corresponding to the different areas mentioned before.

INTRODUCTION

1. The DNA and the central dogma of molecular biology

Almost every single molecule involved in any biological function on living beings has its source in the deoxyribonucleic acid (DNA). DNA encodes most of the instructions required in the development and functioning of an organism. The basic structure of DNA is based on a combination of nucleic acids or nucleotides (Fig. 1). These nucleic acids are small molecules with a high capacity to form stable polymer chains due their atomic structure. The structure of a nucleotide is defined by a monosaccharide a nitrogenous base and a phosphate group that allows the pairing between bases. In practical terms, and focusing only in DNA, what differences a nucleotide from another is its nitrogenous base, which can be adenine (A), citosine (C), guanine (G) or thymine (T). The sequential combination of these 4 letters (ACGT) defines the genomic sequence of an organism.

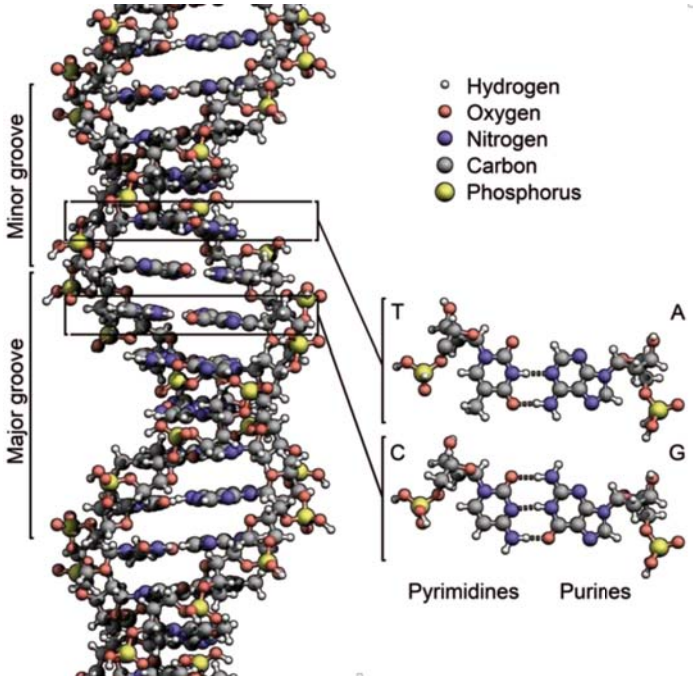


Fig. 1. The atomic structure of nucleotides and DNA

The structure of DNA is characterized by a double helix made up of two strands of nucleotides (Watson and Crick, 1953). Usually, when we refer to the genomic sequence of an organism, we

think only in the sequence of bases that are in one of these strands, because one chain is always complementary reverse to the other. This complementarity occurs because, under normal conditions, an adenine (A) always forms a link with complementary thymine (T), and cytosine (C) always binds a guanine (G). This is due to the biochemical structure of the bases, where a purine (A, G) binds preferentially with pyrimidine (C, T) due to the number of hydrogen bonds they can form (AT has 2 bonds and CG has 3).

The spaces between strands are called grooves and the fact that the geometry of the strands are not symmetric causes differences in their size, giving name to the major and minor groove. Most of the protein-DNA interactions happen in the major groove due to its increased accessibility (Pabo and Sauer, 1984).

Although the genomic sequence defines a guideline for what will eventually materialize in various metabolic processes, the DNA sequence has no functional ability. The conversion from the sequence of nucleotides to a functional unit is explained by the central dogma of molecular biology (Crick, 1970)(Fig. 2).

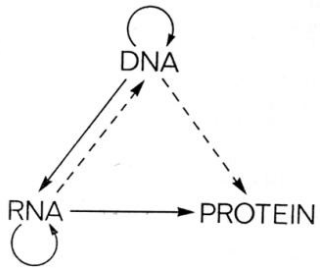


Fig. 2. Central dogma of molecular biology, revised version. In dashed lines, the relations added in the 1970 (Crick, 1970)

Until 1970, a reductionist view of this dogma defined a linear process where DNA is translated to ribonucleic acid (RNA) and RNA is later transcribed to proteins. However, in 1970 the model was reformulated accounting a more complex relationship. RNA role goes beyond being a mere messenger between DNA and the cellular machinery responsible for protein synthesis. RNA can replicate itself without the need of a DNA template, can influence DNA replication and can interfere with other RNA molecules, changing the rules of gene regulation (Hannon, 2002).

2. Nucleosomes, chromatin and epigenetic modifications

The DNA in a cell is folded into chromosomes, highly convoluted and compacted by a factor as much as of 10,000 (Kornberg and Lorch, 1999). As the length of DNA in a chromosome is typically more than 1000 times the diameter of the cell, the DNA must be highly curved in its compacted form inside the cell. The fundamental repeating unit of DNA organization is a structure called the nucleosome.

Nucleosome is the fundamental repeat unit of the eukaryotic DNA. In its canonical form, it consists of 147 DNA base pairs (bp) wrapped 1.6 turns around an octamer of histone proteins with a linker space among adjacent nucleosomes of about 20bp (Richmond and Davey, 2003). This structure is often referred as *beads-on-a-string* for its organization similar to a pearl necklace. In a higher level of compaction, this chain is super-coiled giving place to a blend of DNA and proteins called chromatin, which forms the chromosomes (Fig. 3).

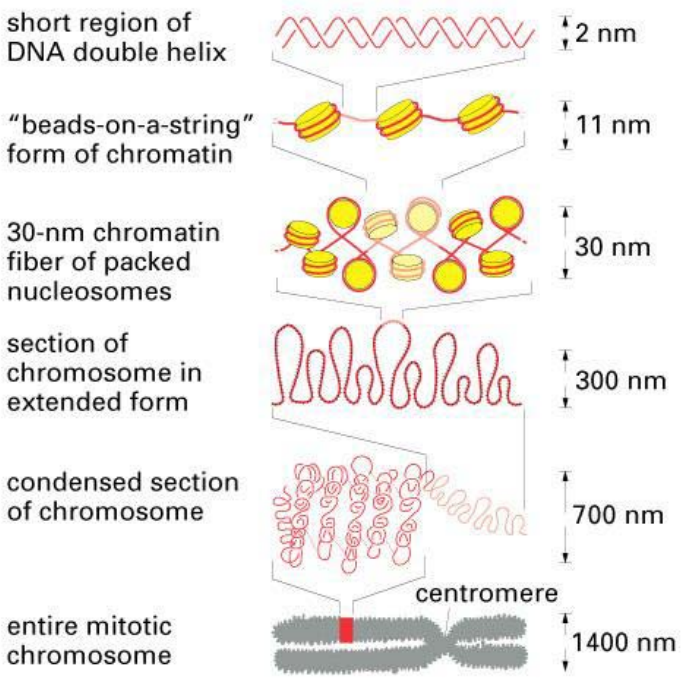


Fig. 3. The different levels of chromatin packaging, from DNA double helix to chromosomes. (Alberts *et al.*, 2004) (Adapted)

The canonical nucleosome core particle contains 2 copies each of the histones H2A, H2B, H3 and H4, but other histone variants are also common in the genome and are associated with altered transcriptional states (Kamakaka and Biggins, 2005; Jiang and Pugh, 2009). In higher eukaryotes, as humans, an additional linker histone, H1, is located at the bounds of the binding regions of DNA with the histone octamer.

The molecular structure of histones is highly conserved among different species and consists of 3 alpha helices which allow the polymerization, exposing their N-terminals to the solvent. These N-terminals, called histone tails, are the target for many post translational covalent modifications such as methylation, acetylation or phosphorylation and show a direct relationship with the transcription and the chromatin state (Ernst and Kellis, 2010), regulating the nearby genes through the so called *histone code* (Jenuwein and Allis, 2001) (Fig. 4).

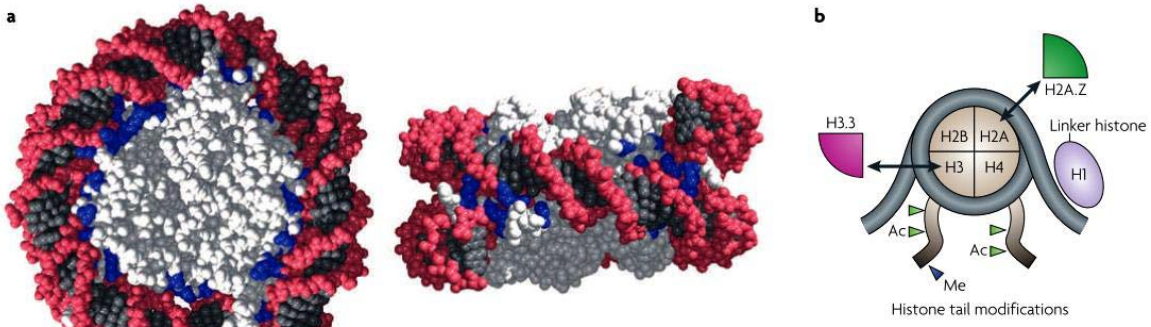


Fig. 4. a) Representation of the atomic structure of a nucleosome, front and side view. DNA is represented as a black rod with a pink backbone; histones appear in white. Blue areas are histone residues in contact with the DNA helix. b) Schematic representation of a nucleosome, with histone variants and modifications of the histone tails. (Jiang and Pugh, 2009) (Adapted)

The first X-ray derived crystal structure of nucleosome at 2.8Å resolution was obtained in 1997 (Luger *et al.*, 1997), despite lower-resolution models were obtained previously revealing some properties of the nucleosome core particle as its disk-shaped structure and the left-handed DNA (Finch *et al.*, 1977; Richmond *et al.*, 1984). A later work provided an improved structure at 1.9Å of resolution by accounting one more basepair step in the crystal (passing from 146bp to 147bp), allowing a clearer electron density and accuracy for the histone protein and DNA atomic coordinates (Davey *et al.*, 2002; Richmond and Davey, 2003)(Fig. 5). In 2010, a new structure was published accounting for a different DNA sequence, but it had a lower resolution (Vasudevan *et al.*, 2010).

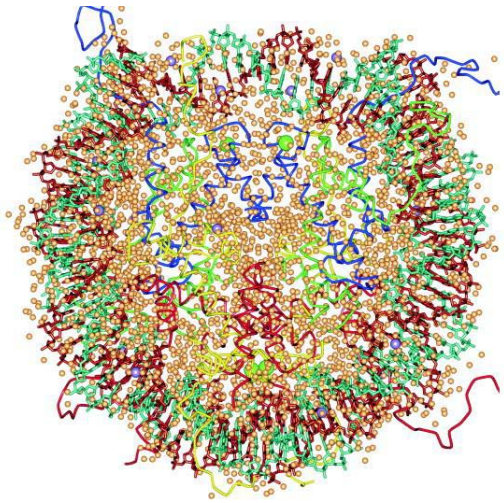


Fig. 5. Atomic representation of half nucleosome core particle plus water molecules for the structure with Protein Data Bank code 1KX5. (Davey *et al.*, 2002) (Adapted)

High-resolution structures of nucleosome core particle revealed different conformation in the wrapping DNA than the one observed in the oligonucleotides on other DNA-protein complexes (Richmond and Davey, 2003). Nucleosomal DNA base-pair-step curvature is higher than the one expected for an ideal B-DNA superhelix of 1.67 turns, affecting the base-pair stretching and kinking the minor groove of the DNA in favor of the major groove. These unusual DNA conformations have implications in the sequence-dependent protein recognition and also in the nucleosome positioning, which can be favored or disfavored in function of the underlying sequence (Segal *et al.*, 2006; Olson and Zhurkin, 2011).

The regulatory role of the histone code and nucleotide modifications, together with its ability to be transmitted from one cell generation to another, points to the fact that not all the genetic inheritance can be explained in terms of DNA primary sequence, hence the name of epigenetics (*above genetics*) (Waddington, 1959).

The field of epigenetics caused a revolution in the way we thought cell differentiation. Few years ago, it was thought that the process by which pluripotent cells sharing the same DNA are differentiated during the development was due to sustainable changes in the gene expression controlled by developmental transcription factors. Nowadays, there are increasing evidences that chromatin regulators have different roles in the cell fate (Bird, 2007). Some of these regulators include changes in the chromatin packaging and accessibility (Bell *et al.*, 2011), DNA modifications (such as methyla-

tion of CpG bases) (Smith and Meissner, 2013), post-translational modification of the histone tails (Chen and Dent, 2014), the incorporation of histone variants (Skene and Henikoff, 2013), ATP-dependent chromatin remodeling (Chen and Dent, 2014) or non-coding RNA mediated pathways (Guan *et al.*, 2013).

Epigenetic factors are not only involved in cell fate but also in disease. Different types of cancers, metabolic or behavioral disorders, cardiovascular, autoimmune or neurodegenerative diseases have been linked with epigenetic marks (Esteller and Herman, 2002; Egger *et al.*, 2004; Portela and Esteller, 2010). The need of a better understanding of those marks and the interplay between genetics, epigenetics and the associated phenotypes promoted huge efforts of the scientific community to deeply explore and map the epigenome landscape in large scale projects such as ENCODE (Dunham *et al.*, 2012), NIH Roadmap Epigenomics (Bernstein *et al.*, 2010) or BLUEPRINT (Adams *et al.*, 2012), among others (Rivera and Ren, 2013).

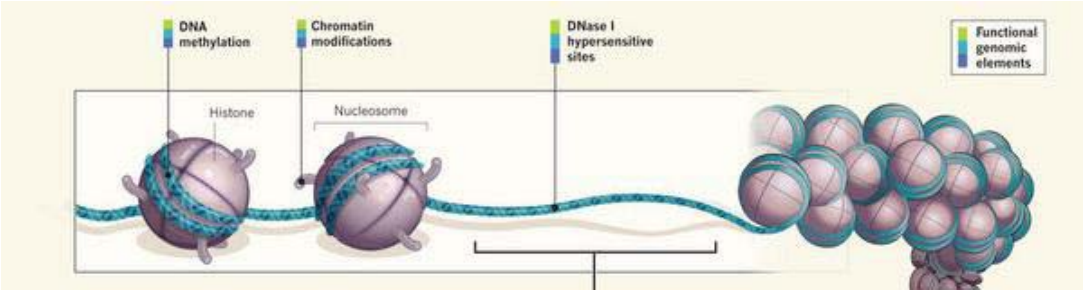


Fig. 6 Summary of main structural modifications of chromatin. (Ecker *et al.*, 2012) (Adapted)

Nucleosomes, as the basic unit of compaction of chromatin, can be understood as an epigenetic regulator as the wrapping of the DNA around the histone cores restricts the accessibility of the underlying sequence, playing a broad regulatory role in the gene expression. Large regions of supercoiled nucleosomes, called heterochromatin, are not accessible to other proteins while other regions lightly packed, called euchromatin, are enriched in active genes (Grewal and Moazed, 2003). Apart from this large-scale effect of DNA compaction, nucleosomes can also act as local-level regulators, for example preventing the binding of Transcription Factors (TF) to the DNA or favoring it through modifications of the histone tails (Portela and Esteller, 2010; Zentner and Henikoff, 2013) (Fig. 6) . In this case, the role of ATP-dependent molecular machinery, such as chromatin remodelers, plays a key role to reposition or eject nucleosomes guaranteeing access to the underlying DNA (Yen *et al.*, 2012).

Apart from nucleosomes, another main player in epigenetics we will introduce in this thesis is DNA methylation. DNA methylation is the addition of a methyl group to a cytosine nucleotide in a CpG basepair and this is a key element in different biological processes such as gene silencing, imprinting or X chromosome inactivation among others (Law and Jacobsen, 2010). The addition or removal of this methyl group is possible due to the action of methyltransferases and demethylases, which act in a context-specific manner allowing the inheritance of methylation patterns along different generations of cells (Jones, 2012; Jones and Liang, 2009).

DNA Methylation is strongly linked with cell differentiation and development (Cedar and Bergman, 2012; Smith and Meissner, 2013). In embryos, DNA methylation is erased and methylation patterns are acquired in a tissue-specific manner in later stages of the development. Also, disruption in methylation patterns are found in many diseases and cancer types (Bergman and Cedar, 2013; Varley *et al.*, 2013; Portela and Esteller, 2010).

A certain consensus exists on the fact that the methylation of CpG islands (regions with an higher accumulation of CpG bases) is a steady mechanism of gene silencing whereas histone modification would provide a more flexible and dynamic mechanism of gene regulation (Cedar and Bergman, 2009), the relationship between DNA methylation and nucleosome positioning is still not well understood. Experimental evidences of a preferential co-localization of nucleosomes in methylated DNA sequences were found in Arabidopsis thaliana (Chodavarapu *et al.*, 2010). These results were further confirmed by *in vitro* experiments showing that a fraction of DNA sequences, preferentially associated with exons, have an increased affinity for histone octamers after the methylation of certain CpG steps (Collings *et al.*, 2013). In this work, authors also found that unmethylated CpG islands near transcription start sites became enriched in nucleosomes upon methylation. However, other works point in the opposite direction. Genome-wide mapping of nucleosomes and DNA methylation in individual molecules, found that methylation profiles are anti-correlated with nucleosome occupancy, being the linker-DNA a preferential target for methylation (Kelly *et al.*, 2012). A recent work featuring clustered methylation patterns in lower eukaryotes also provides evidence that nucleosomes are positioned preferentially between methylated CpG clusters (Huff and Zilberman, 2014). A recent study also shows a relationship between nucleosomal DNA and demethylation: induced DNA demethylation evicts nucleosomes from previously methylated CpG islands, providing a new insight of a synergetic effect between nucleosome positioning and DNA methylation (Portela *et al.*, 2013).

A theoretical approach to the effects of DNA methylation on nucleosome positioning will be presented in the article *Impact of Methylation on the Physical Properties of DNA* (page 49)

3. Nucleosome organization

The placement of the nucleosomes in the genome is not stochastic, but preferred positions rich or poor in nucleosomes are conserved in different cells (Kornberg and Stryer, 1988). In the last decade, thanks to the development of first tilling arrays, and later next generation sequencing, the positioning of nucleosomes in different species has been deeply studied.

Initially, low resolution genome-wide nucleosome maps were done in yeast (Yuan *et al.*, 2005), but quickly evolved in quality and complexity including organisms as worm (Valouev *et al.*, 2008), fly (Mavrich, Jiang, *et al.*, 2008) and recently human (Dunham *et al.*, 2012). Some features are shared along all these maps, as a nucleosome depletion upstream the Transcription Start Site (TSS) and near the Transcription Termination Sites (TTS) (Jiang and Pugh, 2009). A particular nucleosome pattern is also found in other loci, such as exon starts and polyadenylation sites (Tilgner *et al.*, 2009; Spies *et al.*, 2009). The Nucleosome Free Region (NFR) in the TSS is surrounded by two neighboring nucleosomes with similar positioning among different cells, which receive the numeration of -1 (upstream) and +1 (downstream) relative to the NFR (Fig. 7).

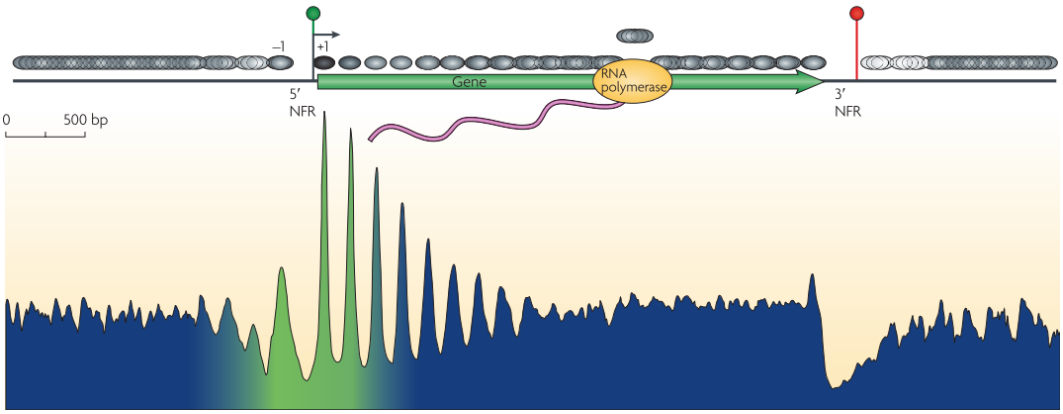


Fig. 7. Genome average distribution of nucleosomes in yeast. Top track represents an array of nucleosome callings (grey ovals) around a gene body (green arrow). NFR at the TSS (green dot) and TTS (red dot) show a lack of nucleosomes. At the bottom, the associated nucleosome coverage. Large accumulations of phased nucleosomes show narrow and large peaks, meanwhile they become a noisy signal when the synchrony is lost. (Jiang and Pugh, 2009) (Adapted)

When a nucleosome shows a similar phasing along a certain number of cells, it shows a narrow peak of coverage and is commonly called well-positioned. In contrast, a nucleosome with a high deviation in the positioning is called fuzzy. Well-positioned nucleosomes are usually found in the start and the end of the gene, but the degree of fuzziness increases as we move away from those regions (Fig. 7).

Obviously, the presence of these common signals motivated many groups to study a general, predictable pattern of nucleosome positioning and/or occupancy. Nucleosome positioning is the position of the histone core respect the wrapping sequence; meanwhile occupancy is the coverage or depletion of nucleosomes in a given locus. Computational analysis of *in vitro* nucleosomal DNA sequences revealed common sequence-dependent preferences for rotational settings. This yielded to the suggestion that a genomic sequence can regulate the nucleosome positioning (Segal *et al.*, 2006; Kaplan *et al.*, 2009). Later studies of nucleosome maps obtained using Next Generation Sequencing (NGS) and inclusion of different organisms in the analysis suggested that this code has a low predictive power of the *in vivo* positioning/occupancy (Stein *et al.*, 2010; Valouev *et al.*, 2008). Different approaches for prediction of nucleosome positioning from primary sequence have been proposed during the last years, but any attempt to confidently predict *in vivo* occupancy fails at some point due to the interplay of *cis*-regulatory elements (Segal and Widom, 2009; Struhl and Segal, 2013). However, the results obtained in *in vitro* experiments certainly point to a preference for a periodic positioning of the nucleosomes respect the underlying sequence, probably due to intrinsic mechanics of the DNA bending (Trifonov, 2010b; Cui and Zhurkin, 2010; Struhl and Segal, 2013). DNA mechanics and their relationship with nucleosome positioning are reviewed in the next section, *DNA physical properties and chromatin modeling*. A detailed presentation of different methods for nucleosome prediction can be found in the section *Nucleosome prediction and modeling* (page 22).

4. DNA physical properties and chromatin modeling

Regarding DNA modeling, many studies about the DNA structural and physical properties were performed since the discovery of the double-helix structure (Watson and Crick, 1953). In this vast field, the introduction of computational simulations of the DNA was a big step forward which allowed

structural biologists to increase dramatically the level of detail of their modeling using Molecular Dynamics (MD) (Tidor *et al.*, 1983; Levitt, 1983).

DNA is a double helix polymer of nucleic acids which base pair together. Depending on the geometry of this double helix, we can find different forms of DNA being B-DNA is the most abundant form in cells (Richmond and Davey, 2003). Other forms of DNA, such as A-DNA or Z-DNA differ significantly from canonical B-DNA in geometry and dimensions (Fig. 8).

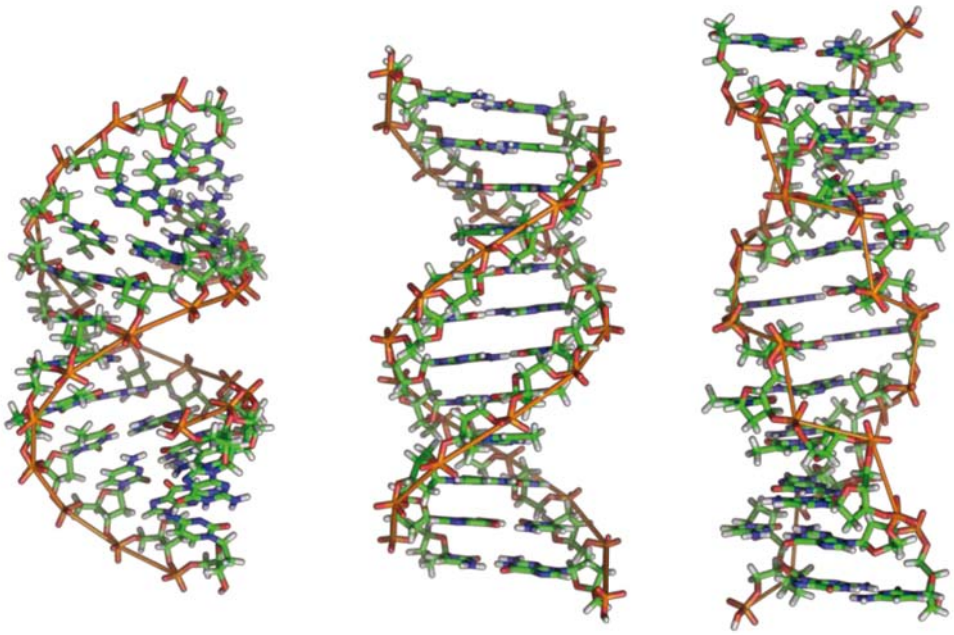


Fig. 8. Different geometries of DNA. From right to left: A-DNA, B-DNA, Z-DNA. (author: R. Wheeler)

If we focus in the base pair geometry, in the tridimensional space we can describe 3 translations and 3 rotations for either the bases facing each other and the basepairs in relationship with the neighboring basepair (Fig. 9).

The measure and the modeling of these helical parameters enables the mesoscopic evaluation of DNA mechanics. In mesoscopic terms, the behavior of a molecule of DNA can be described as a semi-elastic polymer where every basepair has a preferred rotational and translational configuration (Dickerson *et al.*, 1989). With the help of MD simulations and a force field specifically designed for simulations with nucleic acids (Pérez, Marchán, *et al.*, 2007), this computer-derived data can be aggregated into an elastic description of the different degrees of freedom of the DNA at the base-

pair resolution level (Lankas *et al.*, 2003; Morozov *et al.*, 2009; Battistini *et al.*, 2010; Yamasaki *et al.*, 2009).

Different groups used this mesoscopic modeling of DNA to evaluate and predict the structural changes and constrains in the nucleosomal DNA (Olson, 1998; Morozov *et al.*, 2009; Trifonov, 2010a). In general, all these methods take into account different conformational signals from the DNA sequence and compare them with the experimental-derived geometries. A detailed view of our specific implementation of this model and the relationship with other topics featured in this thesis will be introduced later in the results section *Experimental impact of theoretical DNA mechanics* (page 32).

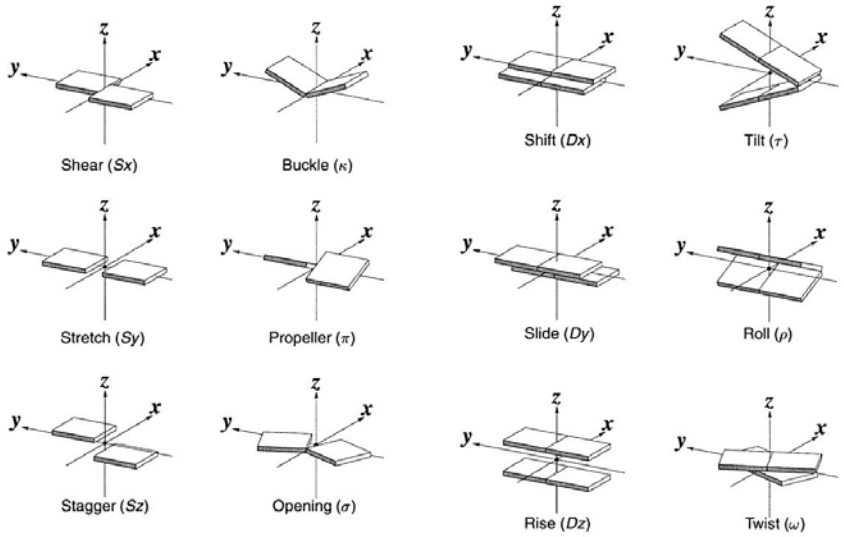


Fig. 9 Helical parameters of the DNA basepair (left) and basepair step (right) (Lu and Olson, 2003)

Despite DNA physical properties are sequence-dependent, the associated description of the basepair steps and their neighboring bases provides a biophysical meaning of the DNA interactions with a richer variability than those provided only by the primary sequence (Lavery *et al.*, 2010a). This description of the physical properties are not only a pure theoretical description useful for structural analysis, but also correlated with regulatory regions *in vivo* (Goñi *et al.*, 2007; Tilgner *et al.*, 2009). In the articles *Physical properties of naked DNA influence nucleosome positioning and correlate with transcription start and termination sites in yeast* (page 35) and *Unraveling the hidden DNA structural/physical code provides novel insights on promoter location* (page 60), we will present two works linking DNA physical properties and regulatory regions.

5. Genome wide nucleosome maps: experimental approaches

During the last years, different methods such as DNase I sensitivity (Lutter, 1989) or chromatin immunoprecipitation of histones cross-linked with DNA (Solomon *et al.*, 1988) were used for the study of chromatin structure. Recently, new methods for the conformational study of chromatin provided new insights in the three-dimensional organization of chromosomes (Baù *et al.*, 2011). Despite the ability of these methods to provide a coarse-grained view of regions with open or histone occupied DNA, the digestion of chromatin with Micrococcal Nuclease (MNase) allowed the retrieval of mononucleosome fragments, enabling a basepair resolution for the study of nucleosome positioning (Teng *et al.*, 2001).

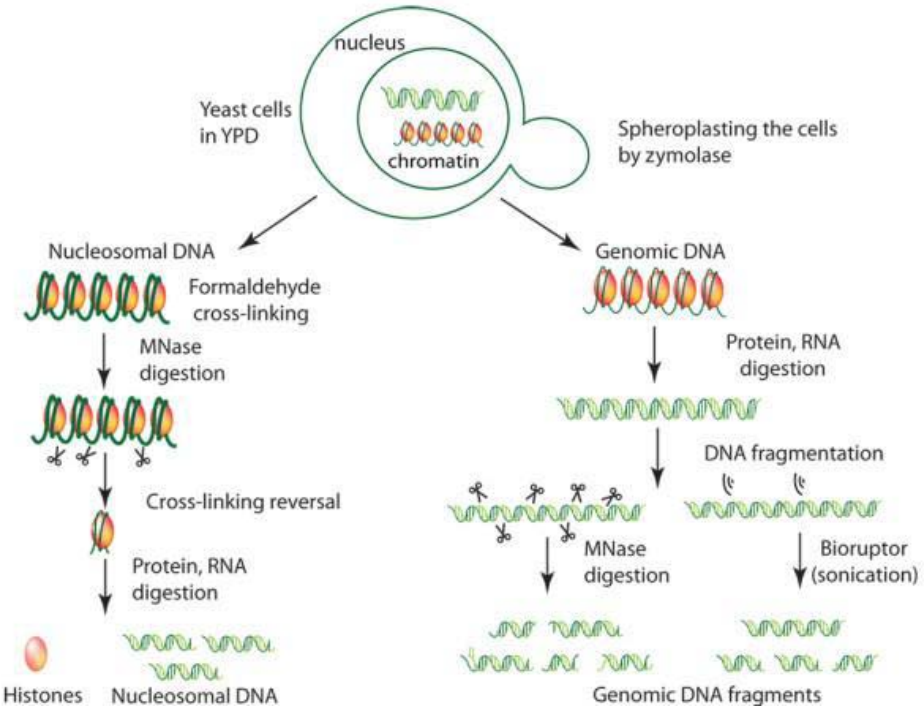


Fig. 10. Schema of experiments for obtaining nucleosomal (chromatin) and genomic (naked) DNA. For nucleosomal DNA sample (left), proteins are cross-linked to their binding sites in vivo with formaldehyde (bold green) and fragmented with MNase before protein digestion. For naked DNA samples (right), proteins and RNA are removed and DNA is fragmented either with MNase or sound waves. The obtained fragments can be then be sequenced or hybridized on microarrays. (Deniz *et al.*, 2011) (Adapted)

A schema of this protocol is presented in Fig. 10. Briefly, genetic material is extracted from cultured cells by disrupting the cell wall. Then, DNA and histones are cross-linked using formaldehyde and

digested with MNase to obtain mononucleosomes. In the last step, proteins and RNA are removed and remaining DNA is loaded in an agarose gel. Electrophoresis should reveal a band over 150bp, which should be the nucleosomal DNA ready to be sequenced or loaded into a microarray (Fig. 11).

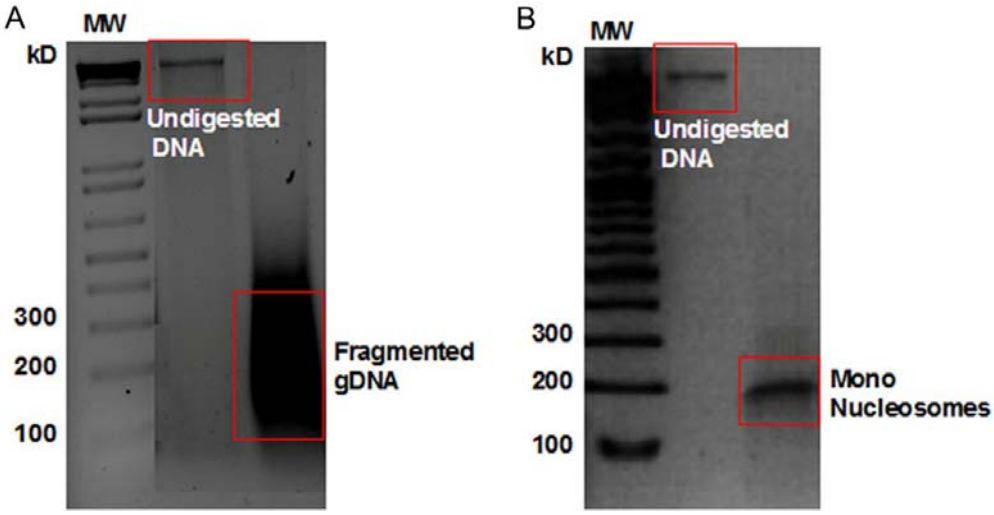


Fig. 11 Native 2% agarose gels showing the genomic (A) and chromatin DNA (B) digestion products before and after MNase treatment, respectively. Fragment sizes were estimated according to standard DNA molecular markers (MW). (Deniz *et al.*, 2011)

Despite MNase has a known preference for specific sequences (Flick *et al.*, 1986; Hörz and Altenburger, 1981), this enzyme has been widely used in the obtention of nucleosome maps (Segal *et al.*, 2006; Lee *et al.*, 2007; Yuan *et al.*, 2005; Schones *et al.*, 2008). Since few years ago, there was still controversy about the potential biases of MNase-digested DNA for the study of nucleosome positions. Meanwhile some authors claimed that MNase bias affects observable nucleosomes maps (Chung *et al.*, 2010a), other authors defended this effect is negligible (Allan *et al.*, 2012). We will present our own conclusions regarding this point in the article *Physical properties of naked DNA influence nucleosome positioning and correlate with transcription start and termination sites in yeast* (page 35) and in the summary of the collaboration realized during a short-stay in the laboratory of Jason Lieb in the University of North Carolina (page 105). Although MNase-Seq (MNase digestion of chromatin followed by next generation sequencing) is nowadays the most used technique for the mapping of nucleosomes, there are alternative methods for obtaining high resolution maps of nucleosome positioning using different types of chemical cleavage or engineered histones (Flaus *et al.*, 1996; Brogaard *et al.*, 2012; Nikitina *et al.*, 2013). All these alternative methods try to

avoid the biases of sequence specificity or different digestion of nucleosome molecule ends due to the use of MNase and could redefine the nucleosome mapping protocol in the future.

The first genome-wide study of nucleosome positioning was performed by partially scanning the genome of Saccharomyces Cerevisiae (baker's yeast) using low-resolution tiling microarrays (Yuan *et al.*, 2005). With this technology, genomic (histone free) DNA and nucleosomal DNA are differentially labeled with fluorescent beads and hybridized to a microarray which contains a region of the genome covered by different tiled oligonucleotides. This first study covered some parts of the yeast genome with a shifting of 20bp between the probes. Two years later, the development of the technology allowed the development of a high density array covering almost the totality of the yeast genome, yielding to the first high-resolution nucleosome map in yeast (Lee *et al.*, 2007). In this case, the tiling between probes was only of 4bp.

Despite tiling arrays were the first high-throughput technology used in genome-wide maps, with the emergence of NGS technologies the resolution of the nucleosome maps increased to one basepair resolution (Shivaswamy *et al.*, 2008; Valouev *et al.*, 2008). Although its name, Next Generation Sequencing usually refers to 2nd generation sequencing, referring to those technologies that appeared after chain-termination related methods (Sanger and Coulson, 1975) and enabled the study of large genomes in few hours with a good coverage. Nowadays, the 3rd generation of sequencers are being developed, with the distinctive trait of working with single DNA/RNA molecules instead of requiring a Polymerase Chain Reaction (PCR) amplification of sequences (Liu *et al.*, 2012).

The different experimental nucleosome maps obtained and related analysis performed during this thesis will be presented in the articles *Physical properties of naked DNA influence nucleosome positioning and correlate with transcription start and termination sites in yeast* (page 35), *Nucleosome architecture and plasticity along cell cycle* (page 74) and *Fuzziness and noise in nucleosomal architecture* (page 78), as well as two collaborations done with external labs (page 92 and 94).

Worth to say that advances in sequencing techniques and the increased resolution and read-depth does not avoid some issues in the analysis of nucleosome maps. The first one regards the use of a population of cells for the DNA extraction. Despite analyzing cells under similar conditions, we can find subpopulations of cells in different phases of cell growth, having differential gene expression levels and chromatin structure. In Fig. 12 we can see the effect of having nucleosomal DNA from cells in different states in a given locus. While some nucleosomes show a perfect phasing (often

nearby strongly regulated regions; see Results sections), others are delocalized showing a lower and fuzzier coverage occupancy. We should be especially careful regarding this point when we average the nucleosomal architecture of different genes. In this case, the big picture obtained from common signals could hide a non-uniform nucleosome positioning landscape.

The synchronized patterns observed over important gene populations near the TSS yielded to the definition of *nucleosomal barriers* for those stable nucleosome-depleted regions. A nucleosome pushed towards this barrier will be well-positioned, but the synchrony of adjacent nucleosomes will be following a statistical positioning model (Mavrich, Ioshikhes, *et al.*, 2008). Furthermore, the fuzziness of a nucleosome map can be accentuated with the presence of nucleosomes with different widths either caused by MNase digestions with different efficiency or non-canonical histone configurations (Zlatanova *et al.*, 2009). In the article *Fuzziness and noise in nucleosomal architecture* (page 74), we measured different sources of noise as well as determinants of nucleosome positioning by sequencing samples of biological replicates under different controlled conditions.

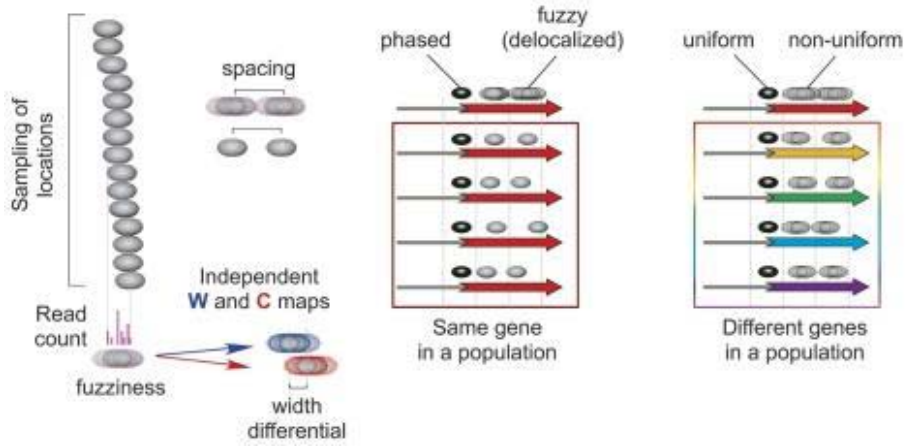


Fig. 12. The effect of populations in nucleosome positions. On the left, the description of the fuzziness from an array of nucleosomes in a given locus coming from different cells. In the middle and right panels, the effect of nucleosome phasing across different cells or different genes inside a same cell. (Mavrich, Ioshikhes, *et al.*, 2008) (Adapted)

Sequencing has become more accessible financially and has become the first choice for genomic studies, including nucleosome maps. The relatively low-cost and an increased resolution compared to tiling arrays caused the last being considered a deprecated technique for obtaining nucleosome maps. Different sequencing technologies are currently available and many more will come in the next years. Among the sequencing technologies existing today, the most commonly used are Ge-

nome Analyzer/HiSeq from Illumina, SOLiD from Life Sciences and 454 GS FLX from Roche, each one of them with different advantages and disadvantages (Liu *et al.*, 2012). Illumina technology is by far the most used in the large sequencing facilities. This is due to its wide range of applications and the lower cost per base (Liu *et al.*, 2012). Recently, new so-called *benchtop* or *personal* sequencers offer low-cost sequencing machines useful for small scale experiments, very suitable for personalized medicine or sequencing of small organisms (Loman *et al.*, 2012).

The 2nd Generation Sequencing has dramatically evolved in the last years lowering costs and enhancing the performance and reliability, but those technologies are not exempt of problems and require an exhaustive quality control check (Patel and Jain, 2012; White *et al.*, 1993). Maybe the biggest issue of NGS is the PCR-amplification in order to increase the amount of DNA/RNA molecules, enabling its recognition. This causes some sequences to be overrepresented (Aird *et al.*, 2011). The combination of this with natural repeated sequences hinders the effective counting of reads. This is specially problematic with short sequences, which are usually overamplified compared with longer ones (Dabney and Meyer, 2012). Also, related with the PCR amplification, CG-rich regions are underrepresented, especially in Illumina experiments (Benjamini and Speed, 2012). Other common problems refer to the correction of reads mapping to opposite strands and the presence of repetitive regions of the genome, but those are effects of the biological organization of the genome and not a pure technological problem (Park, 2009; Zhang *et al.*, 2008). Until new advances in the sequencing techniques will be able to solve these issues, the only way to deal with this kind of problems lies in the computational analysis. Regarding this point, we collaborated with the IRB Biostatistical unit in the development of a package for the quality assessment of high-throughput sequencing experiments and their correction. This work will be presented in the article *htSeqTools: high-throughput sequencing quality control, processing and visualization in R* (page 115)

6. Genome wide nucleosome maps: computational approaches

6.1. Bioinformatic tools

The complexity of the new experimental techniques require a computer in many of the wet-lab procedures, from the analysis of DNA fragments lengths to the base-calling pipeline required by sequencing machines. Besides this class of applications, here we will talk about the tools which enable a computer-driven analysis of the samples.

In the processing of MNase-Seq experiments we find two steps: preparation of the data (pre-analysis) and manipulation of it in order to extract meaningful information (downstream analysis).

In the pre-analysis a series of standard genomic tools are required, including pipelines for the base-calling of sequencing experiments (the programs which convert the electrical signals of the sequencer to a nucleotide sequence) or the microarray tools which measure the fluorescence from scanned images and assign them to a specific sequence/coordinate in the genome. For nucleosome maps, both sequencing and microarray technologies require a reference genome where to map the measured nucleosomes occupancies. In the case of microarrays, the genomic sequences are attached physically on the surface of the different probes on the array. In the case of sequencing, short-reads (sequences of few tens of bases) require mapping process to the reference genome using sequence alignment (Marco-Sola *et al.*, 2012; Langmead *et al.*, 2009).

In early nucleosome maps made with Tiling Arrays, the low-resolution and noisy nature of this technology forced the researchers to transform this data into a non-overlapping set of regions which could be classified as nucleosomal or linker DNA (Yuan *et al.*, 2005; Lee *et al.*, 2007). This was done in the first reference works with Hidden Markov Models (HMM) (Fig. 13).

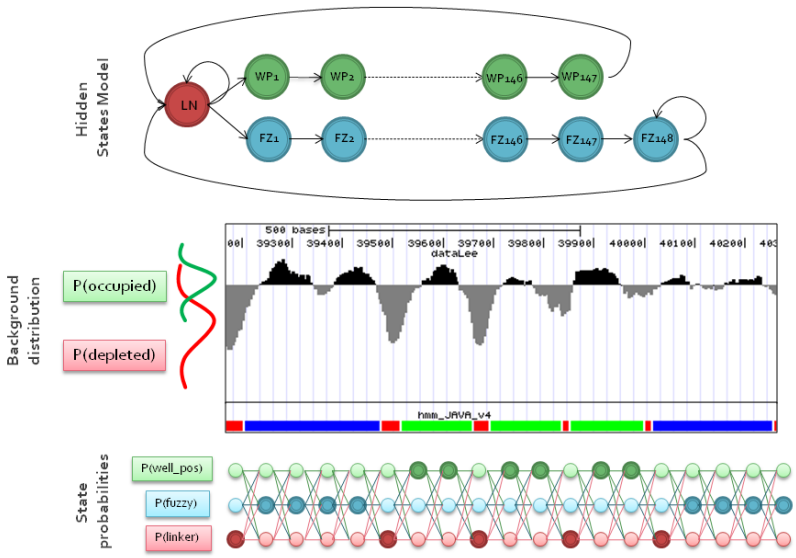


Fig. 13 Custom implementation of a HMM for tiling array data. On top, the persistence transition model with 3 states: linker DNA (LN), well-positioned nucleosomes (WP) and Fuzzy nucleosomes (FZ). At the bottom, the states network measure the most probable state, represented at the bottom part of the middle plot (red=linker, green=well-positioned nucleosome, blue=fuzzy nucleosome)

The process of passing from a numeric, continuous occupancy data to a discrete range of sections representing non-overlapping nucleosomes is a process known as *nucleosome calling*, and often that nucleosome-related publications use both the occupancy values and the nucleosome calls in their analysis (Segal *et al.*, 2006; Lee *et al.*, 2007).

Despite HMM are powerful, flexible, can be used both for microarrays and sequencing data and have been used extensively in different implementations (Xi *et al.*, 2010; Sun *et al.*, 2009; Kuan *et al.*, 2009; Yuan *et al.*, 2005; Yassour *et al.*, 2008), they have few problems associated. The main one is the computational performance of this kind of algorithms, based on a backtracking programming model. This means that we need to scan the whole dataset forwards and backwards before making conclusions. This problem grows in function of the size of the genome and the resolution of the data and it makes this approach unfeasible for 1bp resolution datasets. Another problem with HMMs is the high subjectivity of the modeling. Either transition model or background distributions need to be carefully adjusted in order to retrieve the expected results, but even then, the possible outputs of the algorithm are constrained due to this modeling. Apart from HMM, other methods of nucleosome calling based on Bayesian networks or multi-layer methods were released by this time (Di Gesù *et al.*, 2009; Chen *et al.*, 2010), but still there was an existing need for fast and reliable callers which motivated the development of new approaches in the last years (Becker *et al.*, 2013; Zhang *et al.*, 2012; Polishko *et al.*, 2012; Nellore *et al.*, 2012). In this point we can include our own tool, *nucleR: a package for non-parametric nucleosome positioning* (page 109).

With the technical evolution of computer and sequencing techniques, the level of details of the analysis of nucleosome positions has moved from the global description of non-overlapping nucleosome calls to the analysis of dynamic sub-populations of single nucleosome configurations (Schöpflin *et al.*, 2013; Chen *et al.*, 2013). Accordingly, we developed a new method of analysis of nucleosome dynamics at single read level. Details can be found in the publication *Dynamic analysis of nucleosome positioning at read level* (page 125).

To finish with the analysis of bioinformatics methods, I would like to mention the R/Bioconductor framework (Gentleman *et al.*, 2004), which allows a highly customizable and efficient integrative analysis of biological data. In fact, all the developed libraries presented in this thesis in the section *Algorithmic and computational methods* (page 106) and few others developed for internal use but not released are programmed over this language.

Linked with our extensive use of R, we felt in different occasions the need of integrating our analysis with pre-existent database such as Ensembl, UCSC Genome Browser or Uniprot and the lack of a native way to query these public datasets within R. To solve this problem we developed a simple and convenient connector for servers using Distributed Annotation System (DAS) protocol. This is presented in the publication *DASiR: Programmatic data retrieval from DAS servers in R* (page 118).

6.2. Nucleosome positioning prediction and modeling

Besides the bioinformatic tools required for the analysis of high-throughput experiments, the prediction and modeling of nucleosome positions has been in the last decades a hot topic either for theoreticians and experimentalists working with chromatin. After observing certain sequence periodicities in the chicken nucleosome cores (Satchwell *et al.*, 1986) and assuming that nucleosomes locations are not random (Kornberg and Stryer, 1988), different methods and models have been proposed for the prediction and description of nucleosome positioning signals which we summarize below.

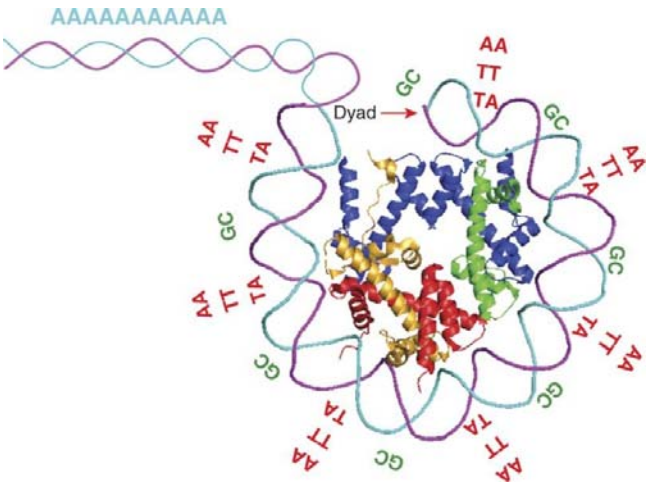


Fig. 14. Lateral view of a histone tetramer showing the proposed basepair periodicity which favors the rotational positioning of a nucleosome. Despite the fact that alternate patterns of AT-GC are favorable for nucleosome positioning, poly(A·T) elements are reluctant to nucleosome formation and found enriched in NFR. (Struhl and Segal, 2013) (Adapted)

One of the most relevant scientists in this area was Jonathan Widom, which carried out a SELEX experiment quantifying the affinity of different synthetic random DNA molecules for histone octamers (Lowary and Widom, 1998). In a later work, a ~10 bp periodicity in TA dinucleotide step

were identified in the sequences showing the highest affinity (Thåström *et al.*, 1999) (Fig. 14). In spite of this, those high-affinity molecules were non-natural and affinity for genomic sequences was shown less specific than the one observed in the in vitro experiment. Further work in this direction provided which probably has been the most impactful article in this area so far, claiming the discovery of *a genomic code for nucleosome positioning* (sic) (Segal *et al.*, 2006). The model proposed in this article was later revised by the same authors and two additional models for predicting nucleosome positioning from sequence were proposed (Field *et al.*, 2008; Kaplan *et al.*, 2010)

In parallel, other groups also presented their own methods and conclusions regarding nucleosome positioning prediction using different techniques sharing a starting point in the statistical inference of patterns from experimental or in vitro nucleosome maps (Peckham *et al.*, 2007; Gupta *et al.*, 2008; Tillo and Hughes, 2009; Chung and Vingron, 2009; Ioshikhes *et al.*, 2011). Some of those methods not only focus in the patterns favouring nucleosome formation but also disfavouring it. For example, poly(dA·dT) tracks or TGGA repeats are proved to be sequences reluctant to adopt a nucleosome conformation (Cao *et al.*, 1998; Wu and Li, 2010)

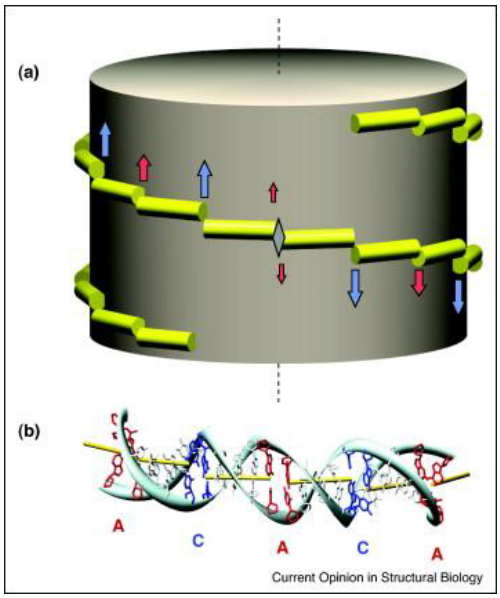


Fig. 15. a) Simplified representation of stair-case like superhelical pathway of nucleosomal DNA basepairs (yellow rods) adapting an ideal cylindrical histone octamer (grey core). In the center, nucleosome dyad is marked as a diamond and 4 base pairs steps are shown in details in panel b). (Olson and Zhurkin, 2011) (Adapted)

Another approach derived from the sequence analysis is trying to understand the biophysical or mechanistical basis by which certain sequences are more affine to nucleosomes than others. Different groups have integrated the structural knowledge of DNA and nucleosome core crystal structures with the sequence-dependent preferences explained above. Most of those models try to explain the affinity of given sequences according their intrinsic flexibility or underlying ability to be deformed adopting an ideal nucleosome structure (Olson, 1998), global bendability or conformational measures (Gabdank *et al.*, 2010; Cui and Zhurkin, 2010; Tilgner *et al.*, 2009) or specific patterns of certain helical parameters (Balasubramanian *et al.*, 2009; Tolstorukov *et al.*, 2007; Olson and Zhurkin, 2011) (Fig. 15).

Finally, a third group of models are those basing their predictions in biological signals. We find methods based on comparative genomics and evolutive models (Ioshikhes *et al.*, 2006; Tsankov *et al.*, 2011; Möbius *et al.*, 2013), models which propose that just few nucleosomes in the genome have their position determined (either by intrinsic DNA sequences or the effect of in vivo factors, such as chromatin remodelers or transcription factors) and adjacent nucleosomes are phased statistically according those (Mavrich, Ioshikhes, *et al.*, 2008; Teif and Rippe, 2009; Milani *et al.*, 2009; Lubliner and Segal, 2009) or integrative models which take into account all those different signals (Parmar *et al.*, 2013).

In the other side, some authors are skeptic about establishing precise rules for nucleosome positioning based on DNA sequence (Zhang *et al.*, 2009; Stein *et al.*, 2010). A more consensus view states that intrinsic properties of DNA play indeed a key role in nucleosome positioning, but predictive models should also take into account the large impact of in vivo factors in the structure of the chromatin (Segal and Widom, 2009; Struhl and Segal, 2013).

Worth to say that, despite the development of a nucleosome predictor from primary sequence was one of the first projects during this thesis, this is still an ongoing work. The reasons are many, and they include the need of a better understanding of the chromatin organization before developing the model and the fact that we used non-discrete energy potentials to describe and predict *in vivo* nucleosome conformations. Our model for the calculation of DNA deformation energy is presented in the section *Experimental impact of theoretical DNA mechanics* (page 32).

OBJECTIVES

1. Experimental impact of theoretical DNA mechanics

- Investigate the relationship of intrinsically high or low flexible sequences with regulatory regions and *in vivo* nucleosome maps
- Adjust existing mesoscopic models of DNA flexibility to support DNA methylation and genome-wide analysis
- Provide bioinformatics support to the wet lab in the experimental validation of theoretical results

2. Nucleosome organization in vivo

- Process MNase-seq data in order to retrieve our own nucleosome maps
- Perform quality control analysis and study possible biases on MNase-seq experiments
- Study the impact in the chromatin of differential gene expression and DNA replication during the cell cycle in yeast.
- Establish collaborations with strong experimental groups for joint research in chromatin organization and regulation, providing our expertise and computational resources

3. Algorithmic and computational methods

- Develop a predictor of nucleosome positioning based on physical descriptors of DNA.
- Set up a processing pipeline for the analysis of microarray and sequencing experiments, especially nucleosome maps obtained from MNase-seq experiments.
- Review existing methods appropriated for our analysis and develop new ones if required.
- Adapt existing algorithms or methods to take advantage of high performance computing available in the group.

PHD ADVISOR REPORT

Publication:

Flores, O. and Orozco, M. (2011) "*nucleR: a package for non-parametric nucleosome positioning*", Bioinformatics (Oxford, England) 27: 2149–50.

O.Flores is the first author and the main developer of NucleR.

Publication:

Deniz, Ö.*, Flores, O.*, Battistini, F., Pérez, A., Soler-López, 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.

O. Flores is first (co)author of the paper. He performed all the bioinformatic analysis in the paper.

Publication:

Pérez, A., Castellazzi, C. L., Battistini, F., Collinet, K., Flores, O., Deniz, Ö., Ruiz, M. L., Torrents, D., Eritja, R., Soler-López, M., Orozco, M. (2012) "*Impact of Methylation on the Physical Properties of DNA*", Biophysical Journal 102: 2140–8.

O. Flores performed the bioinformatics analysis of the impact of methylation in the organization of the nucleosome architecture in S.Cerevisae.

Publication:

Nadal-Ribelles, M.*, Conde, N.*, Flores, O., González-Vallinas, J., Eyras, E., Orozco, M., de Nadal, E., Posas, F. (2012) "*Hog1 bypasses stress-mediated down-regulation of transcription by RNA polymerase II redistribution and chromatin remodeling*". Genome Biology 13: R106.

O.Flores was the first author of the group in this collaborative project. He took responsibility in analyzing the nucleosomal maps obtained by F.Posas's group.

Publication:

Planet, E., Stephan-Otto, C., Reina, O., Flores, O. and Rossell, D. (2012) "*htSeqTools: high-throughput sequencing quality control, processing and visualization in R*", Bioinformatics (Oxford, England) 28: 589–90.

O.Flores contributed in this project as part of a laboratory rotation. He helped in the improvement of the performance and execution times of the library.

Publication:

Duran, E., Djebali, S., Gonzalez, S., Flores, O., Mercader, J. M., Guigó, R., Torrents, D. Soler-López, M., Orozco, M. (2013) "*Unravelling the hidden DNA structural/physical code provides novel insights on promoter location*", Nucleic Acids Research 41(15):7220-30

O.Flores was involved in the bioinformatic analysis of the data and in the processing of ProStar predictions.

Publication:

Flores, O. and Mantsoki, A. (2013) "DASiR: Programmatic data retrieval from DAS servers in R", Bioconductor (http://www.bioconductor.org/packages/release/bioc/html/DASiR.html)

O.Flores is the first author and the main developer of DASiR package.

Publication:

Flores, O.*, Deniz, Ö.*, Soler-López, M., Orozco, M. (2014) "*Fuzziness and noise in nucleosomal architecture*", Nucl. Acids Res. (early access) doi: 10.1093/nar/gku165.

O.Flores is first (co)author. He was the responsible of all the bioinformatic analysis in the project.

Publication:

Flores, O. and Orozco, M. "Dynamic analysis of nucleosome positioning at read level" (in preparation)

O.Flores is the first author of this paper and the main developer of the dynamic model.

Publication:

Deniz, Ö.*, Flores, O.*, Soler-López, M., Orozco, M. "*Nucleosome architecture and plasticity during the cell cycle*", (in preparation)

O. Flores is first (co)author of the paper. He performed all the bioinformatic analysis in the paper.

Ö. Deniz is co-author of the following articles as part of her thesis: *Physical properties of naked DNA influence nucleosome positioning and correlate with transcription start and termination sites in yeast, Fuzziness and noise in nucleosomal architecture* and *Nucleosome architecture and plasticity during the cell cycle*. Ö. Deniz was responsible of designing and performing the experiments in these articles and her task was complementary and equally relevant to the contribution made by O. Flores.

The advisor:

Modesto Orozco López

PUBLICATIONS

1. Experimental impact of theoretical DNA mechanics

My group, Molecular Modeling and Bioinformatics (MMB), has been intensively studying the physical properties of the nucleic acids with Molecular Dynamics over the last decade, with high impact contributions to this field (Noy *et al.*, 2005; Pérez, Luque, *et al.*, 2007; Pérez *et al.*, 2008; Orozco *et al.*, 2008; Pérez, Marchán, *et al.*, 2007)

In the previous years to this thesis, the relationship of those intrinsic physical properties of the nucleic acids have been linked with regulatory regions and *in vivo* functions, both in our group (Goñi *et al.*, 2008) and in others (Cairns, 2009; Tilgner *et al.*, 2009). Above all, the most discussed topic in this area has been the relationship between the intrinsic curvature and bend ability of the DNA and its associated potential to form nucleosomes (Trifonov, 2010b). Despite the extensive number of works in this field, DNA-histone interactions are still a source of controversy (De Santis and Scipioni, 2013). The main debate is about the predictive potential of pure physical-mathematical models or the knowledge-based models (i.e., using data mining and machine learning to include experimental information to the model) to reproduce experimental-annotated nucleosomal sequences.

In general, published models are precise in the prediction of good or bad *in vitro* nucleosome reconstitution, but their predictive power fall *in vivo* maps (Kaplan *et al.*, 2009; Morozov *et al.*, 2009; Cui and Zhurkin, 2010). Anyway, despite a precise prediction of the nucleosome positioning may be not possible, we proved in the past that promoter detection from primary sequence is possible accounting only for physical descriptors of the DNA (Goñi *et al.*, 2007). In a similar way, DNA physical parameters were correlated with other genomic regions of regulatory interest, such as exon boundaries (Tilgner *et al.*, 2009).

With the creation of the Experimental Bioinformatics Laboratory (EBL), associated to the MMB, we had a first-hand source of experimental data. This data can be experimental validations of theoretical models, required in the benchmark of our computational research. Also, the collaboration between MMB and EBL can be given in the opposite direction, when experimental designs require a

bioinformatics support in order to process the experiments, perform data mining or retrieve information in a massive way.

When I joined MMB, my first task was helping in the development of a nucleosome predictor from primary DNA sequence using a new set of refined DNA descriptors applied in a simple harmonic model (Olson, 1998). The proposed model is based on the calculation of free energy required for a given nucleosome-long sequence to adapt the histone core shape. The relative difference in terms of geometrical positioning of the different base pair step helical parameters in the three-dimensional space (Fig. 9, page 14) is measured between an average DNA conformation and the conformation provided by the DNA in a reference nucleosome crystal structure. With a simple harmonic oscillator, the force constant associated with this shift in the geometry provides the final increment of energy, which formulation (1) will be:

$$\Delta E = \sum_{par} \sum_{bp} K(x - x_0)^2 \qquad (1)$$

where *ΔE* is the sum for every helical parameter *par* and every base pair step *bp* of a given sequence of the harmonic potential with force constant *K* and displacement $x - x_0$, being *x* the value of the base pair in the reference crystal structure and x_0 its value in the equilibrium state (Lankas *et al.*, 2003). Our model uses K and x_0 values derived *ab initio* from our own MD simulations using a modified force field specially designed for nucleic acids simulations (Pérez, Marchán, *et al.*, 2007). *x* values were obtained from a smoothed average of the available nucleosome structures in the Protein Data Bank (PDB). This model was later refined to account for neighboring effects of different base pairs (Lavery *et al.*, 2010a). This will be the model which will be used, unless otherwise stated, in the rest of this document and associated articles when referring to *deformation energy*.

In a first version of the model, we had the hypothesis that some specific pairs of <*par, bp*> could have a different contribution to the final deformation energy value, for example, according to their position in the minor or major groove of the DNA. Our objective was to retrieve a set of nucleosomal and naked DNA sequences to train a matrix of weights with an expectation maximization method. This motivated the implementation of a Hidden Markov Model (see page 20), a new method for nucleosome calling (nucleR, page 109) and the making of our own nucleosome map using microarrays and high-throughput sequencing. In the process of obtaining those *in vivo* good and bad positioning sequences, we found that the effect of MNase digestion on the control samples

(naked DNA) was bigger than expected. This motivated the research presented in the article *Physical properties of naked DNA influence nucleosome positioning and correlate with transcription start and termination sites in yeast* (page 35). Worth to say that finally we discarded the trained/weighted model in favor of the pure *ab initio* model, which yields to good results without requiring a model fitted to a specific genome.

In the frame of the collaboration between MMB and EBL, I provided bioinformatics support in a new study about the effects of methylation in the physical properties of the DNA. The aim of this work was to obtain and study the properties of a new set of parameters similar to the ones explained above in this section but which account the base pair step methyl-CpG and the neighboring ones. This study is presented in the publication *Impact of Methylation on the Physical Properties of DNA* (page 49). In the last point of this chapter, I also collaborated with other members of MMB-EBL in the experimental validation of the method for detection of regulatory regions using the physical descriptors of the DNA published by MMB, ProStar (Goñi *et al.*, 2007). Details are presented in *Unraveling the hidden DNA structural/physical code provides novel insights on promoter location* (page 60).

1.1. Physical properties of naked DNA influence nucleosome positioning and correlate with transcription start and termination sites in yeast

This project starts with the need for training our theoretical predictor of the deformation energy using an experimental set of nucleosomal DNA sequences. Besides the training, this information will allow us to evaluate and improve the performance of the predictor. Different sets of sequences obtained from SELEX experiments were reported in the literature by this time (late 2008 - early 2009), providing us with a limited set of examples for high or a low nucleosome affinity sequences (Thåström *et al.*, 1999; Lowary and Widom, 1998). Despite this, our intention of performing training of more than 876 variables pairs <helical parameter, position>, required a larger set of sequences and high-resolution nucleosome map which were not still available. Therefore, we designed together with EBL an experiment to obtain our own nucleosome map in yeast, consisting of a sample of nucleosomal DNA and naked DNA control as shown in Fig. 10 (page 15).

After the preprocessing of the samples and some preliminary analysis, we found that, consistent with our model, the calculated deformation energy from primary sequence was anti-correlated with the nucleosome occupancy profile (Fig. 16). This means that regions of high energy were indeed nucleosome depleted in the gene boundaries and the phasing of the -2, -1 and +1 nucleosomes relative from the TSS could be explained by this energetic barrier.

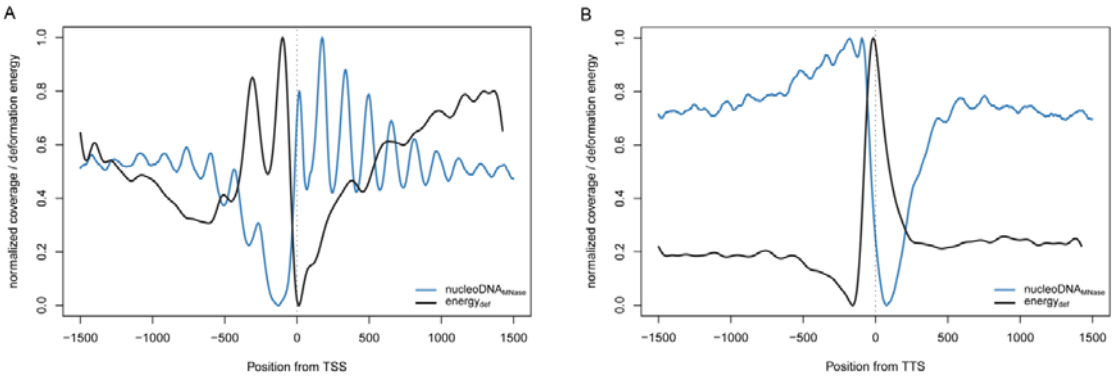


Fig. 16. Experimental nucleosome maps (blue) and calculated deformation energy (black) genome-wide profile around TSS (A) and TTS (B).

Nevertheless, the comparison of our nucleosomal DNA and the naked DNA control showed an abnormal digestion pattern around the TSS. The correlation between the naked DNA and the nucleosomal DNA was higher than expected in the NFR upstream the TSS, warning us about a po-

tential bias of MNase in nucleosome maps. We thought about two possible bias hypotheses: the fact that NFRs could be, indeed, an artifact of a preferential digestion of the enzyme or an artifact caused in the amplification which was affecting both the sample and the control. The study of both MNase cutting sites and digested regions reported similar patterns between both samples but not in a naked DNA fragmented using a physical shredding (sonication) (Fig. 17), discarding the later hypothesis.

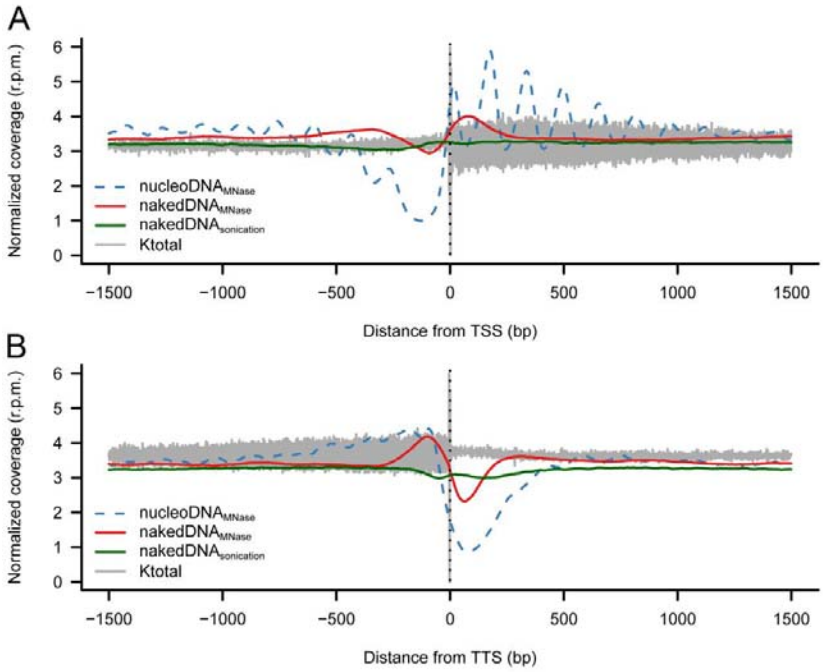


Fig. 17 Profiles of genome-wide nucleosome coverage (mean) around the TSS (A) and TTS (B) in yeast. Blue-dashed line represents the nucleosomal DNA, red line the naked DNA and dark green line the sonicated DNA. Ktotal (the aggregation of the different force constants) shows a differential physical characterization in the gene body and the upstream/downstream regions.

The detailed analysis of the physical descriptors in these Low coverage Regions (LRs) of both nucleosomal and naked DNA revealed a local increment of the flexibility near the cutting sites which could explain a certain affinity for the endonuclease activity of MNase (Fig. 18). At the same time, global physical descriptors predicted high deformation energy around LRs, consistent with the fact that those regions were really nucleosome depleted *in vivo* (Fig. 18). Despite similar conclusions were achieved by other groups during the publication stage of this paper (Chung *et al.*, 2010b), our approach proved that potential biases of MNase occur preferentially in nucleosome-depleted areas and link the intrinsic DNA physical properties with *in vivo* nucleosome maps.

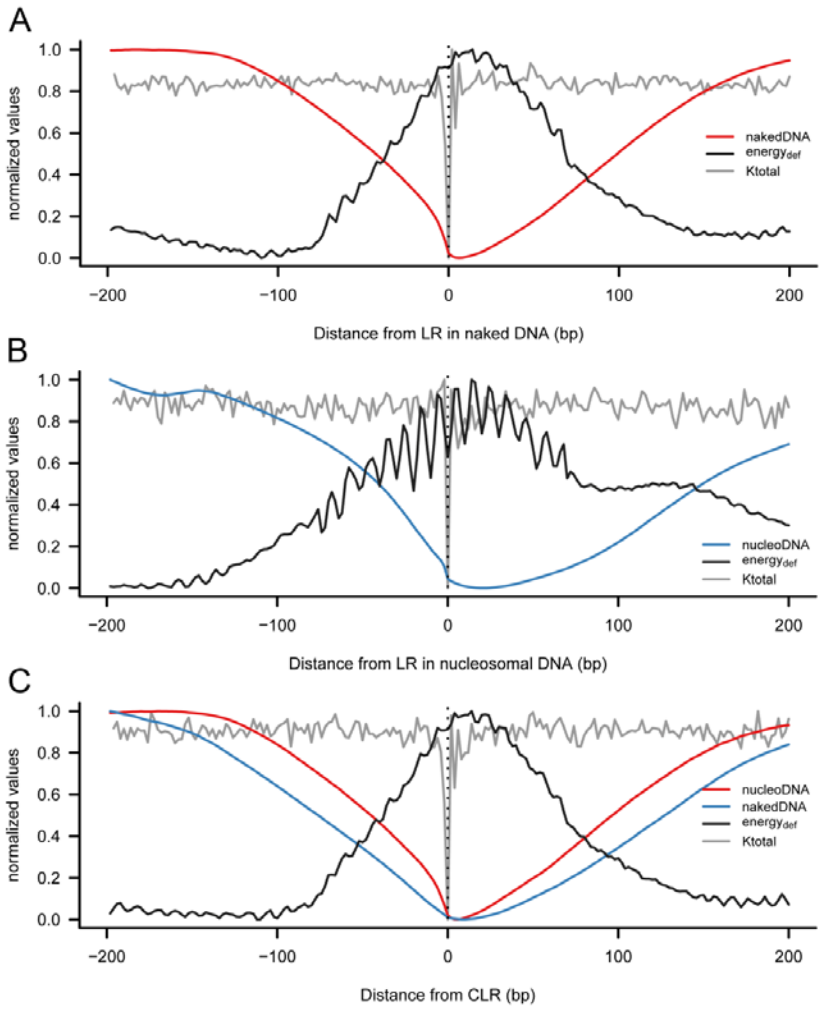


Fig. 18. Digestion profiles around LRs in naked DNA (red)(A), nucleosomal DNA (blue)(B) and Common LRs (CLR)(C). Local DNA tetramer flexibility is represented by *Ktotal* (grey) meanwhile the predicted deformation energy for sequences of 147 bases are represented by $energy_{def}$ (black)

Publication:

Deniz, Ö.*, Flores, O.*, Battistini, F., Pérez, A., Soler-López, 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.

Supplementary material for this article can be found in the page XVI of the Annex 2.

* Equally contributing authors

Deniz et al. BMC Genomics 2011, 12:489
http://www.biomedcentral.com/1471-2164/12/489

BMC Genomics

RESEARCH ARTICLE Open Access

Physical properties of naked DNA influence nucleosome positioning and correlate with transcription start and termination sites in yeast

Özgen Deniz[1†], Oscar Flores[1†], Federica Battistini[1], Alberto Pérez[2], Montserrat Soler-López[1] and Modesto Orozco[1,3,4*]

Abstract

Background: In eukaryotic organisms, DNA is packaged into chromatin structure, where most of DNA is wrapped into nucleosomes. DNA compaction and nucleosome positioning have clear functional implications, since they modulate the accessibility of genomic regions to regulatory proteins. Despite the intensive research effort focused in this area, the rules defining nucleosome positioning and the location of DNA regulatory regions still remain elusive.

Results: Naked (histone-free) and nucleosomal DNA from yeast were digested by micrococcal nuclease (MNase) and sequenced genome-wide. MNase cutting preferences were determined for both naked and nucleosomal DNAs. Integration of their sequencing profiles with DNA conformational descriptors derived from atomistic molecular dynamic simulations enabled us to extract the physical properties of DNA on a genomic scale and to correlate them with chromatin structure and gene regulation. The local structure of DNA around regulatory regions was found to be unusually flexible and to display a unique pattern of nucleosome positioning. *Ab initio* physical descriptors derived from molecular dynamics were used to develop a computational method that accurately predicts nucleosome enriched and depleted regions.

Conclusions: Our experimental and computational analyses jointly demonstrate a clear correlation between sequence-dependent physical properties of naked DNA and regulatory signals in the chromatin structure. These results demonstrate that nucleosome positioning around TSS (Transcription Start Site) and TTS (Transcription Termination Site) (at least in yeast) is strongly dependent on DNA physical properties, which can define a basal regulatory mechanism of gene expression.

Keywords: DNA physical properties, Molecular dynamics, MNase digestion, nucleosome positioning, gene regulation, chromatin structure

Background

Genomic studies mostly provide one-dimensional information encoded in DNA, but we cannot ignore the fact that in eukaryotic organisms, DNA is packaged into chromatin structure, where DNA folds to a global compaction of at least 10^4 [1]. Genome homeostatic histone concentration ensures most of DNA to be wrapped into nucleosomes (~75-90%) [2], which are structural units of 145-147 base pairs (bp) long, where the interaction with regulatory proteins is severely handicapped. Nucleosomes are separated from each other by short linkers (around 20 bp long in yeast) where site-specific recognition by proteins is easier. Therefore, DNA compaction has clear functional implications, since it modulates the accessibility of genomic regions to regulatory proteins. Indeed, a close relationship was established between nucleosome positioning and important regulatory signals [3], such as proximal promoters [4,5] and splicing sites [6]. Further evidence on the connection between three-dimensional

* Correspondence: modesto.orozco@irbbarcelona.org
† Contributed equally
[1]Institute for Research in Biomedicine and Barcelona Supercomputing Center Joint Research Program on Computational Biology. Baldiri i Reixac 10. Barcelona 08028. Spain
Full list of author information is available at the end of the article

chromatin structure and function was obtained from genome-wide analysis of chromatin DNase I degradation profiles, which revealed a cross-link between DNase I hypersensitive sites and regulatory regions [7-9].

DNA underlying sequence has long been considered to be an important contributor to nucleosome assembly [10-13]. Crystal structures of nucleosome core particles revealed a lack of direct readout mechanisms between histones and DNA bases (the so-called base readout) [14-17] which led to the postulate that histone-DNA direct interactions are not the major determinant of nucleosome positioning [18]. Accordingly, the DNA relative affinities for nucleosome formation (e.g. high-affinity Widom601 sequence) [19] should be based on an indirect readout mechanism, where the ability of a given DNA sequence to be deformed would account for the nucleosome assembly preferences [20-24]. Nevertheless, to which extent nucleosome positioning *in vivo* is really dictated by the DNA sequence is still an issue of strong discussion [25-27].

Our group and others have provided indirect evidence highlighting the connection between DNA physical properties and chromatin organization [28-30]. In particular, we have previously reported theoretical studies showing that human promoters display very unusual stiffness properties [31]. These might affect DNA binding of regulatory proteins, either directly by hampering or favoring complex formation, or indirectly through the modulation of the chromatin structure and hence the DNA accessibility [31]. Here, we have pursued this hypothesis by a genome-wide analysis of conformational properties across yeast naked DNA using micrococcal nuclease (MNase) degradation profiles as an experimental descriptor. We were able to characterize in detail, MNase preferences for naked DNA, extending fractional information derived from small-scale experiments. These preferences (at the tetramer level) correlate with *ab initio* physical descriptors derived from molecular dynamics (MD) simulations of short DNA oligonucleotides [32-35]. This finding confirms that MNase can signal genomic regions with unusual physical properties [36,37]. Very interestingly, MNase-hypersensitive sites in naked DNA are mainly located around TSS and TTS, which supports experimentally our suggestion that those regulatory regions are signaled by physical properties. Moreover, the correlation of genome-wide nucleosome positioning profiles with MD-derived mesoscopic calculations evinces that the main mechanism by which physical properties influence gene regulation is through nucleosome positioning. Altogether, our experimental and computational integrative analysis demonstrates a clear relationship between sequence-dependent structural properties of naked DNA, accessible from first-principles simulations, and regulatory signals in chromatin structure.

Results and Discussion

Preferential MNase cut sites

Yeast DNA fragments were prepared and sequenced following the experimental approach described in Figure 1. The analysis of our whole genome sequencing experiments, containing more than 75 million short fragments, provided a fully converged description of the MNase sequence preferences for cutting naked and nucleosomal DNAs (Table 1). As suggested from previous small-scale experiments [38], we indeed observed that in naked DNA, the enzyme preferentially cuts tetramers with a central d(A-T) step, but without the requirement of flanking dC or dG bases suggested by low-scale experiments. The high-cutting susceptibility for d(CATA)·d(TATG) tetramers found in mouse satellite DNA [39] was also detected in our massive experiments, although these tetramers were not the most predominant cutting sites. On the other hand, tetramers resistant to MNase cleavage were very diverse, except for the presence of a central purine-purine dinucleotide step (Additional File 1: Table S1). Overall, MNase displayed quite strong sequence preferences in naked DNA (up to a factor of 200) (Table 1) that could not be simply ascribed to experimental artifacts, given the fact that control experiments where DNA was fragmented by sonication did not show any marked variation in genome-wide profile (Additional File 1: Figure S1). It is noteworthy to mention that MNase resistant tetramers were different between naked and nucleosomal DNA samples, which demonstrate that the nucleosome structure protects specifically certain sequences from MNase degradation. Conversely, we found a good agreement in the preferred cutting sites between naked and nucleosomal DNAs (Table 2). This suggests that tetramer signals that are directing the first MNase cut in chromatin are intrinsic to naked DNA.

Preferential MNase degraded regions

Upon an initial endonucleotic cleavage, MNase displays an exonuclease activity that continues with the degradation of DNA [40], leading to digested areas that we identified as low coverage regions (LRs) in our sequencing experiments (see Methods). Tetramer composition along naked DNA LRs was different from the one observed in the cutting sites, suggesting that the degradation of a particular fragment does not only depend on the existence of cleavage sites in its vicinity, but also on the differential sequence preferences of endo- and exo-nuclease activities. For example, d(AAAA·TTTT) was the most abundant tetramer in naked DNA LRs, nearly four times more frequent than expected ($p < 10^{-8}$), while the same tetramer was rarely present at primary cutting sites (1/4 than expected, $p < 10^{-7}$, Additional File 1: Table S1). Moreover, the tetramer composition was very similar in both naked and

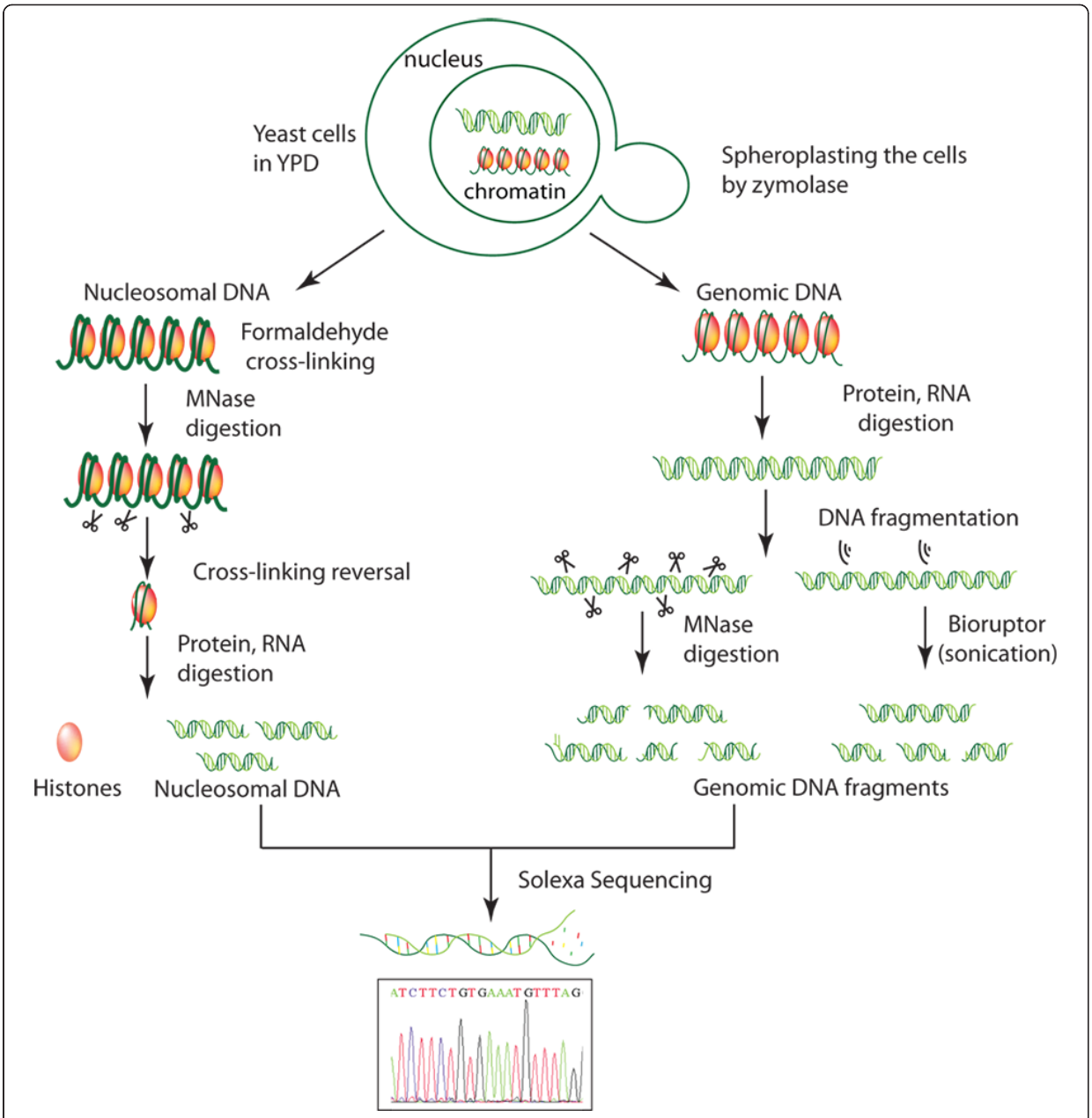


Figure 1 Overview of the experimental procedure of DNA sample preparation for Illumina Sequencing. A population of wild-type *Saccharomyces cerevisiae* was spheroplasted by zymolase. For nucleosomal DNA sample (left), proteins were cross-linked to their binding sites *in vivo* with formaldehyde (bold green) and chromatin was extracted and fragmented with MNase. For naked DNA samples (right), proteins and RNA were removed. Naked DNA was extracted and fragmented either by MNase or Bioruptor system (sonication). All the obtained fragments were sequenced on the Illumina/Solexa Genome Analyzer (GA) IIx.

nucleosomal LRs and in the common low regions (CLRs) (definitions in Additional File 1: Additional Methods) indicating that sequence susceptibility for MNase degradation in nucleosomal DNA was not exclusively dependent on the chromatin structure, but was also related to the intrinsic properties of naked DNA (Table 2).

Low coverage regions and physical properties

The MNase tetramer preferences (Tables 1 and 2) are so diverse that they cannot be explained in terms of direct DNA base reading. Analysis of MD-derived physical properties [32,41] revealed that primary cutting sites are characterized by high flexibility (affecting roll and tilt parameters)

Table 1 Frequency of MNase-preferred tetramers at the cutting sites

Naked DNA	ratio	p-val	Nucleosomal DNA	ratio	p-val
TATA.TATA	13.28	$< 10^{-18}$	CTAG.CTAG	4.07	$< 10^{-18}$
ATAG.CTAT	8.45	$< 10^{-18}$	ATAG.CTAT	3.93	$< 10^{-18}$
CTAA.TTAG	7.90	$< 10^{-18}$	CAAG.CTTG	3.57	$< 10^{-18}$
CTAG.CTAG	6.80	$< 10^{-18}$	CTTA.TAAG	3.52	$< 10^{-18}$
ATTA.TAAT	5.74	$< 10^{-18}$	CATG.CATG	3.42	3.01×10^{-4}
CATA.TATG	5.62	$< 10^{-18}$	CATA.TATG	3.11	$< 10^{-18}$
ATAA.TTAT	5.14	$< 10^{-18}$	CTAA.TTAG	3.00	$< 10^{-18}$
CTTA.TAAG	4.92	$< 10^{-18}$	CTAC.GTAG	2.98	$< 10^{-18}$
TTAA.TTAA	4.64	$< 10^{-18}$	ATTG.CAAT	2.96	$< 10^{-18}$
ATAT.ATAT	4.52	$< 10^{-18}$	AAAG.CTTT	2.82	$< 10^{-18}$
TAAA.TTTA	3.48	$< 10^{-18}$	CTTC.GAAG	2.79	$< 10^{-18}$
ATTG.CAAT	3.25	$< 10^{-18}$	AATG.CATT	2.50	$< 10^{-18}$
GTAA.TTAC	2.64	1.01×10^{-4}	CATC.GATG	2.24	6.03×10^{-4}
ATAC.GTAT	2.39	2.01×10^{-4}	CAAC.GTTG	2.19	10^{-3}
			CAAA.TTTG	2.17	$< 10^{-18}$

Experimentally detected and expected frequency ratios of MNase-preferred tetramers at the cutting sites for naked (left) and nucleosomal (right) DNAs. Displayed tetramers are observed in at least two-fold higher frequency than expected, with a statistically significant difference ($p < 10^{-3}$) (Supplementary Methods). Ratios for d(TAAA)·d(TTTA) and d(GTAA)·d(TTAC) tetramers in nucleosomal DNA are 1.39 ($p < 0.08$) and 1.6 ($p < 0.03$) respectively.

and wide opening in the major groove (high roll values) at the equilibrium geometry (Additional File 1: Figure S2). Furthermore, the total dinucleotide-based stiffness parameter k_{total} (see Methods for definition) unveiled that LRs (in both naked and nucleosomal DNA) are located in regions with large variations in flexibility, where an extremely flexible site is surrounded by stiff motifs (Figure 2 and Additional File 1: Figure S3). Remarkably, the same results were obtained when we considered the parameters fitted to the tetramer level by the ABC consortium [42] confirming the robustness of our conclusions. Dinucleotide and tetranucleotide data (see below and Additional File 1: Additional Methods) are available upon request and are incorporated in our DNAlive webserver (http://mmb.pcb.ub.es/DNAlive),

Nucleosome positioning and gene structure

As previously suggested by other groups ([10,11,43-50]) MNase resistant regions in nucleosomal DNA are mainly concentrated at the beginning of transcribed regions (Figure 3). Whereas very sensitive regions (i.e. LRs) were mostly identified at regulatory regions, either upstream of transcription start sites (TSSs) (Figure 3A) or downstream of transcription termination sites (TTSs) (Figure 3B). Additional differential regions, such as MNase resistant areas upstream of TTSs or downstream of TSSs, were less certain than the major signals mentioned above (Figure 3). Considering that MNase degradation profiles were only dependent on nucleosome positioning [51-55], we could locate more than 33,000 "well positioned" and around 48,000 "fuzzy" nucleosomes along the yeast genome (see

Table 2 Frequency of tetramers in MNase-digested LRs and CLRs

Naked DNA	ratio	p-val	Nucleosomal DNA	ratio	p-val	Common low regions (CLR)	ratio	p-val
AAAA.TTTT	3.87	$< 10^{-18}$	TATA.TATA	4.06	$< 10^{-18}$	AAAA.TTTT	4.48	$< 10^{-18}$
TAAA.TTTA	2.38	$< 10^{-18}$	ATAT.ATAT	3.09	$< 10^{-18}$	TATA.TATA	3.18	$< 10^{-18}$
TATA.TATA	2.38	9.05×10^{-4}	AAAA.TTTT	2.91	$< 10^{-18}$	TAAA.TTTA	2.67	$< 10^{-18}$
AAAT.ATTT	2.16	$< 10^{-18}$	ATAA.TTAT	2.21	$< 10^{-18}$	ATAA.TTAT	2.62	$< 10^{-18}$
ATAA.TTAT	2.13	$< 10^{-18}$	AATA.TATT	2.08	$< 10^{-18}$	ATAT.ATAT	2.57	$< 10^{-18}$
TTAA.TTAA	2.10	7.54×10^{-3}	ATTA.TAAT	1.99	10^{-4}	AATA.TATT	2.43	$< 10^{-18}$
AATA.TATT	2.02	$< 10^{-18}$	TAAA.TTTA	1.84	7.04×10^{-4}	TTAA.TTAA	2.29	3.22×10^{-3}
ATAT.ATAT	2.00	4.62×10^{-3}	AAAT.ATTT	1.62	4.22×10^{-3}	AAAT.ATTT	2.27	$< 10^{-18}$
AATT.AATT	1.84	5.53×10^{-3}				ATTA.TAAT	2.15	$< 10^{-18}$
ATTA.TAAT	1.79	3.62×10^{-3}				AATT.AATT	1.81	1.30×10^{-2}
GAAA.TTTC	1.45	3.44×10^{-2}						

Experimentally detected and expected frequency ratios of different tetramers in MNase-digested LRs for naked (left) and nucleosomal (center) DNAs, and in CLRs (right). Displayed tetramers show a significant enrichment ($p < 0.05$) respect to genome average (Additional File 1: Additional Methods).

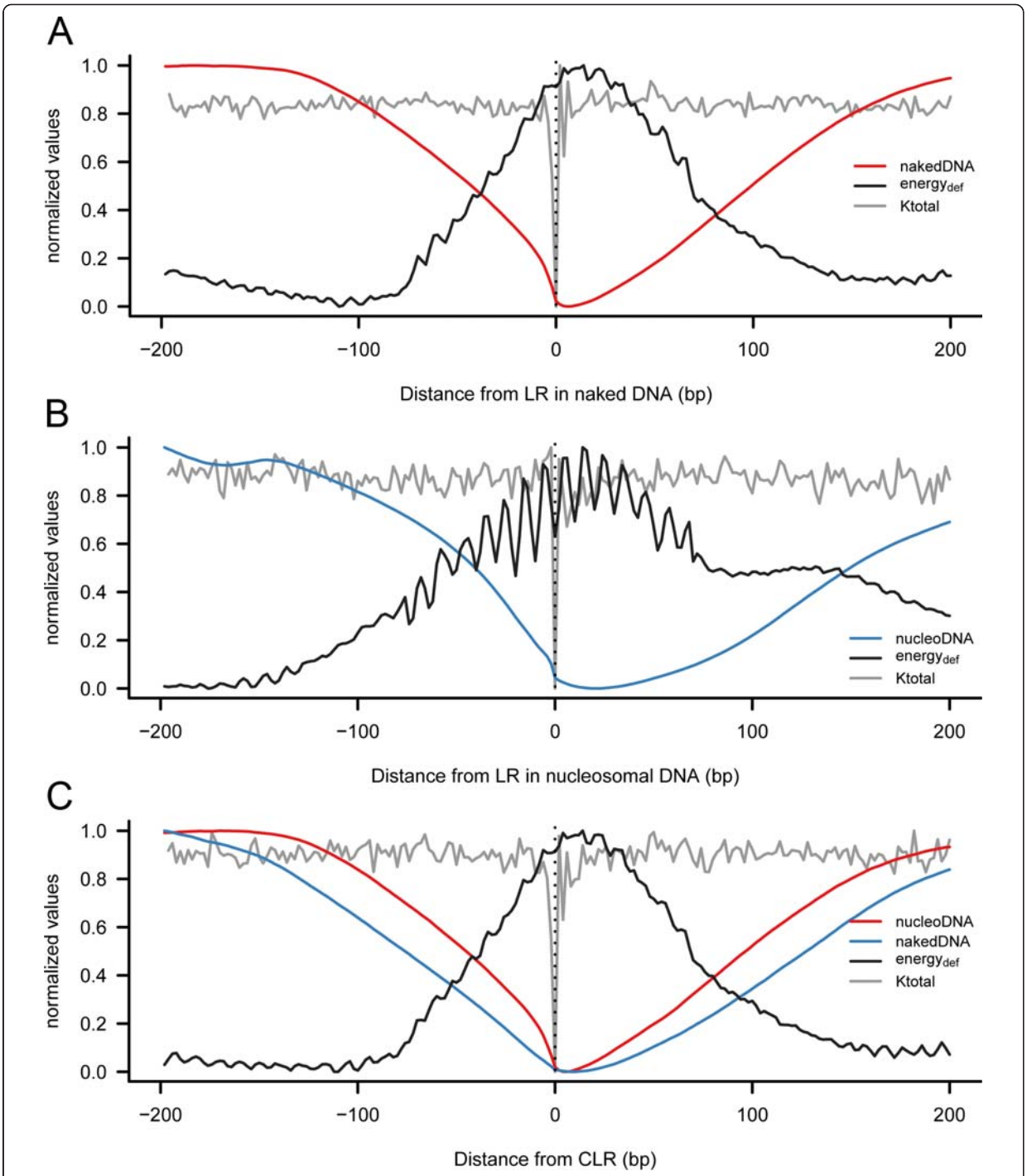


Figure 2 Stiffness, deformation energy and coverage profiles in low coverage regions. Total stiffness parameter (k_{total}), deformation energy and coverage maps were calculated and averaged across all yeast genome, around (A) LRs in naked DNA, (B) LRs in nucleosomal DNA and (C) CLRs in nucleosomal and naked DNA. Deformation energy describes the energetic cost of wrapping a 147 bp DNA fragment into the nucleosome conformation (higher the energy, more the nucleosome location is disfavored). All values are normalized (scaled in the range 0-1) to facilitate analysis and comparisons.

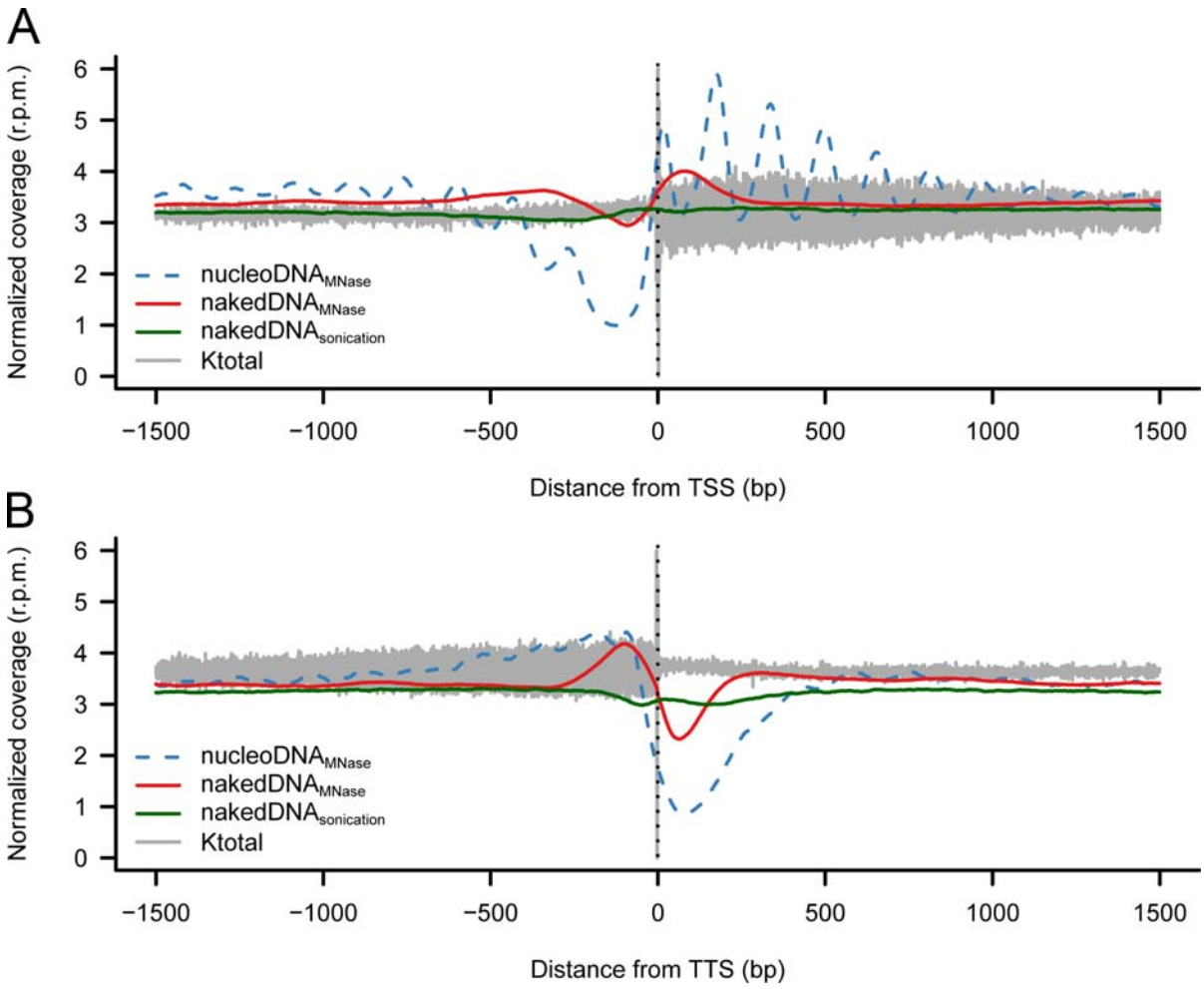


Figure 3 MNase degradation profiles in naked and nucleosomal DNA samples. Coverage maps per base pair were calculated and averaged across all yeast genome around (A) TSSs and (B) TTSs for MNase-digested nucleosomal and naked DNAs and sonicated naked DNA. The averaged total stiffness parameter (k_{total}) profile is shown for comparison.

Methods and Additional File 1: Additional Methods for details). Notwithstanding, the surprising similarity observed between nucleosomal and naked MNase profiles (not detected in the sonication profiles; Figure 3) indicate that nucleosomal degradation profiles might not only reflect nucleosome positioning, but also the intrinsic susceptibility of naked DNA to MNase digestion [56]. This is clearly illustrated in the reduction of nucleosome positioning signals in Additional File 1: Figure S4, when nucleosomal MNase degradation maps are corrected with the naked DNA ones (see Methods). However, Figure 3 clearly demonstrates that strong nucleosome depletion or "well positioned" nucleosome signals, such as upstream of TSS and downstream of TTS, are not affected by the correction of intrinsic MNase susceptibility biases. These observations thus support most of the claims in previously reported nucleosome positioning studies about the connection between nucleosome organization and gene regulation [10,11,43-50] and toned down some recent criticisms about the neglect of the MNase bias.

Physical properties, nucleosome positioning and regulatory regions

The analysis of MD-derived descriptors of naked DNA showed that key genomic regions, such as at TSSs and TTSs, were marked by unusual flexibility properties (Additional File 1: Figure S5) [29]. Since those regions are strongly nucleosome depleted, we hypothesized that unusual physical properties might control nucleosome positioning in those regions, which in turn would affect the DNA accessibility to regulatory proteins and ultimately impact gene regulation. To verify this hypothesis,

we computed the deformation energy required to wrap a DNA sequence around a histone octamer by using a simple elastic energy function based on the MD-derived physical descriptors (see Methods). Figure 2 clearly shows that CLRs, which are nucleosome depleted, correlate with high deformation energy confirming that in these regions it is more difficult to wrap DNA around a nucleosome core. It is interesting to note (Figure 2) that, often, 147 bps regions with high deformation energy contain a high flexible (4 mer) step, indicating that global concepts about the impact of point flexibility on chromatin organization needs to be considered with caution. Overall, results in Figure 2 strongly suggest that the properties that make a DNA segment a good substrate for MNase are also those that avoid DNA wrapping around a nucleosome. In fact, very encouragingly, deformation energies for wrapping a DNA around a nucleosome core particle can accurately predict in vivo nucleosome distribution around TSSs and TTSs in yeast (Additional File 1: Figure S6). These results suggest that, without dismissing the importance of cellular mechanisms for controlling chromatin structure, very important details of the nucleosome organization around TSS and TTS can be rationalized considering physical properties of the naked DNA sequence.

Conclusions

The molecular mechanisms that regulate gene expression in eukaryotic organisms are very diverse and complex. Considering the large amount of basal gene expression in cells, it is difficult to believe that regulation is entirely modulated by specific direct-readout mechanisms, where regulatory proteins would directly interact with DNA through hydrogen bonds in the major/minor grooves and compete with histones [57]. Thus, a combination of direct and indirect readout mechanisms is required to achieve the correct interaction affinity and specificity [58]. Direct mechanism can be very specific, but has implicitly a large energetic cost. Indirect mechanism is obviously less precise, but implies no energy cost for the cell and might be useful in cases where no specific regulation of the gene is needed.

Genome-wide sequencing of MNase treated nucleosomal DNA shows that key regulatory regions such as the start and the end of transcribed sites, which have been traditionally interpreted as nucleosome depletion sites, are actually signaled by a differential pattern of MNase susceptibility in naked DNA. This observation, which could initially raise some concerns, does not contradict previously reported nucleosome maps where MNase degradation was supposed to only reflect nucleosome positioning [10,11,43-50,59-61]. Indeed, nucleosomal degradation profiles corrected with naked DNA data maintained major nucleosome positioning signals, such as nucleosome depletion upstream of TSS or downstream of TTS, thereby supporting previous MNase based nucleosome positioning conclusions [62,63]. Nevertheless, our experiments with nucleosomal and naked DNA suggest caution in the interpretation of nucleosome positioning signals in regions with anomalous MNase degradation profile.

The high correlation of MNase degradation profiles of nucleosomal and naked DNA and with unusual stiffness properties indicates that (without dismissing the importance of the cellular machinery for control of chromatin structure) intrinsic physical properties of naked DNA determine major nucleosome location signals in yeast, especially those at TSS and TTS. This hypothesis is indirectly supported by very recent studies [64], where nucleosome positioning signals are clearly identified after genome-wide nucleosome reconstitution *in vitro*.

Essential regions for gene regulation like TSSs and TTSs are characterized by unusual physical properties that disfavor positioning of nucleosomes and therefore expose DNA to interaction with regulatory proteins. This property of regulatory regions is quite general across the genome. The genes with well-defined CLRs at regulatory regions did not differ from those with more diffuse signals in terms of Gene Ontology analysis [65], promoter architecture, transcription rate or their dependence on regulatory proteins. Accordingly, we can infer that unusual physical properties are perhaps a general property of gene regulatory regions that can confer a basal mechanism of gene regulation. Furthermore, we speculate that additional specific signals were evolutionarily conferred to enable proteins to directly read DNA sequences in those genes that might require a finer regulatory mechanism.

All conclusions drawn here have been derived from the analysis of yeast genome and thus concerns exist whether they can be validated for higher eukaryotes with a different sequence composition at regulatory regions. Therefore, we compared the sequence-dependent physical properties of the *Drosophila melongaster* genome with the high-resolution genomic nucleosome map available [66]. The comparative analysis is shown in Additional File 1: Figure S7, which revealed that coverage and stiffness profiles at TSS are conserved between such distant organisms like yeast and fruit fly [67]. Extension of conclusions to vertebrates is more complex, due to the higher impacts of epigenetic factors. Nevertheless unusual physical properties are also remarkable in human promoters [31]. All these findings prompt us to believe in the general conclusion that nucleosome-depleted and enriched regions are signalled by unusual physical properties, which define the core of an evolutionarily conserved mechanism of gene regulation.

Methods

DNA sample preparation

Both nucleosomal and genomic (histone-free) DNA were isolated from *Saccharomyces cerevisiae* BY4741 strain, (an outline of the experimental procedure is presented in Figure 1, adapted from a previously described procedure) [50]. For nucleosomal DNA preparation, exponentially growing yeast cells were first cross-linked with formaldehyde, spheroplasted with zymolase and finally subjected to a MNase partial digestion to generate core nucleosomes containing DNA fragments of around 147 bp (see Additional File 1: Additional Methods). Agarose gel electrophoresis confirmed that more than 90% of the isolated DNA corresponded to mono-nucleosomal fragments (Additional File 1: Figure S8). Naked DNA was prepared from overnight grown culture by spheroplasting the cells with zymolase and subsequently incubated with SDS and RNase for an efficient protein and RNA depletion. DNA samples were analyzed by fluorometry and UV spectrophotometry to ensure that proteins and RNA were completely removed from the DNA (Additional File 1: Figure S8). The purified DNA was then sheared following two different approaches (MNase digestion and sonication) that yielded fragments of approximately 150 bp in both cases (additional details in Additional File 1: Additional Methods). To guarantee that results were not dependent on MNase concentration, experiments were repeated using two MNase concentrations (0.04 and 0.12 U) (data not shown, but available upon request). The original, the corrected degradation maps and MNase cutting preferences did not show any differeces between the two MNase concentrations. Accordingly in this study only the data obtained with high MNase concentration are reported. These degradation conditions ensure that in nucleosomal DNA experiments only the linker DNA is digested, most of the degraded sample corresponds to mononucleosomes, and integrity of DNA bound to histones is preserved.

DNA sequencing

Cleaved DNA samples were sequenced on the Illumina/Solexa Genome Analyzer (GA) IIx to generate reads of 54 bp length. Data were processed with standard GA base calling pipeline to convert initial raw images into sequences. All sequencing experiments were done in duplicates. Pooled data highly converged, as the reproducibility of individual experiments was very large in all cases (Additional File 1: Additional Methods). Reads are available in Short Read Archive of NCBI under Accession Number SRA030453.

Mapping reads to genome

GA reads were aligned to the *Saccharomyces cerevisae* reference genome using the Bowtie software [68], allowing up to three mismatches per read. Short reads with multiple alignments were mapped to all possible places, thus avoiding the generation of artificial depleted regions. Largely over-represented reads were eliminated to reduce PCR amplification artifacts. Coverage values were calculated for each position on the genome, normalized and converted to reads per million (r.p.m.) (Additional File 1: Additional Methods).

Nucleosome calling and MNase bias correction

Nucleosomes were defined as regions flanking ±74 bases the peaks detected in the coverage maps. Peak detection was performed using a recently published algorithm *nucleR* [69] (Additional File 1: **Additional Methods**). Correction of nucleosomal digestion profiles was done by using the degradation profiles obtained for naked DNA as background (Additional File 1: Additional Methods).

Identification of cut sites and low coverage regions

MNase cut sites were extracted from mapped reads, taking two bases upstream and the two bases downstream of every read end. Low coverage regions (LRs) account for regions were MNase degradation has been especially extensive. Low coverage regions (LRs) were detected in both nucleosomal and naked DNA as genomic segments with non-zero coverage below certain thresholds (Additional File 1: Additional Methods).

Derivation of physical descriptors

Parameters describing the equilibrium geometry and deformability of naked DNA were derived from long atomistic MD simulations of a reduced number of short oligonucleotides (displaying all unique dinucleotide or tetranucleotide steps) in solvent water by using a newly developed force-field [70]. Base pair and base step structures of DNA can be described as a set of three translations (shift, slide and rise) and three rotations (tilt, roll and twist), while the deformability along those directions can be described by their associated stiffness constants (K_i), considering the equilibrium conformation as the origin of energies following the approach suggested by Lankas and others [34,32,42,41]. In brief, the covariance matrix defining the deformability of helical parameters of a given DNA segment (for example a dinucleotide step) is computed from the ensemble of molecular dynamics simulations and inverted to determine 6 × 6 stiffness matrix for each fragment (for example each of the ten unique dinucleotide steps, or the ten dinucleotide steps adapted to all tetramer environments). Pure stiffness constant associated to individual helical deformations (k_{tilt}, k_{roll}, k_{shift}, k_{tilt}, k_{rise} and k_{slide}) are taken from the diagonal of the matrix. K_{total} is obtained as the product of the six pure stiffness constants and gives a rough global estimate of the flexibility of each base pair step (Additional File 1: Additional Methods)

Calculation of nucleosome deformation energy

The energetic cost of wrapping a 147 bp DNA fragment was determined by using an harmonic approach: $E = 0.5\ X^T\ \Theta\ X$; where Θ is the stiffness matrix derived from MD simulations; X (or X^T) is the deformation vector (or its transposed), given by translating a relaxed DNA fiber into the coiled nucleosome core DNA conformation as described for averaging and smoothing of X-ray structures (Additional File 1: Additional Methods). Note that no training is performed and therefore deformation energies are fully *ab initio* descriptors. The scripts used to perform deformation energy calculations are available upon request to the authors.

Additional material

Additional file 1: Additional Methods, Additional Figures and Additional Tables. PDF document with detailed methods and additional results.

List of abbreviations

MNase: micrococcal nuclease; DNase I: DNA nuclease I; MD: molecular dynamics; LR: low coverage region; CLR: common low coverage region; TSS: transcription start site; TTS: transcription termination site; RNase: RNA nuclease.

Acknowledgements

We thank D. Rossell and E. Planet for their support in pre-processing the high-throughput sequencing data and F. Azorín for assistance in the experimental assays. This work was supported by the Spanish Ministry of Science and Innovation (BIO2009-10964 and Consolider E-Science), Instituto de Salud Carlos III (INB-Genoma España and COMBIOMED RETICS) and Fundación Marcelino Botín. AP is an EMBO fellowship holder (ALTF 1107). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author details

[1]Institute for Research in Biomedicine and Barcelona Supercomputing Center Joint Research Program on Computational Biology. Baldiri i Reixac 10. Barcelona 08028. Spain. [2]Laufer center for physical and quantitative Biology, Stony Brook University, Stony Brook, NY 11794, USA. [3]Department of Biochemistry and Molecular Biology. University of Barcelona. Avinguda Diagonal. Barcelona 08028. Spain. [4]Instituto Nacional de Bioinformática. Parc Científic de Barcelona. Baldiri i Reixac 10. Barcelona 08028. Spain.

Authors' contributions

The authors have made the following declarations about their contributions: MO had the idea and make the general planning of the study. Conceived and designed the experiments: OD MS MO. Performed the experiments: OD. Analyzed the data: OF FB AP MO. All authors contributed to the writing of the manuscript. All authors read and approved the final manuscript.

Competing financial interests

The author(s) declare that they have no competing interests

Received: 21 June 2011 Accepted: 7 October 2011
Published: 7 October 2011

References

1. Gilbert N, Boyle S, Fiegler H, Woodfine K, Carter NP, Bickmore WA: **Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers.** *Cell* 2004, **118**:555-566.
2. Holde Kvan, Zlatanova J: **Unusual DNA structures, chromatin and transcription.** *BioEssays : news and reviews in molecular, cellular and developmental biology* 1994, **16**:59-68.
3. Jiang C, Pugh BF: **Nucleosome positioning and gene regulation: advances through genomics.** *Nature reviews Genetics* 2009, **10**:161-72.
4. Lodhi N, Ranjan A, Singh M, Srivastava R, Singh SP, Chaturvedi CP, Ansari SA, Sawant SV, Tuli R: **Interactions between upstream and core promoter sequences determine gene expression and nucleosome positioning in tobacco PR-1a promoter.** *Biochimica et biophysica acta* 2008, **1779**:634-44.
5. Choi JK, Kim Y-J: **Intrinsic variability of gene expression encoded in nucleosome positioning sequences.** *Nature genetics* 2009, **41**:498-503.
6. Schwartz S, Ast G: **Chromatin density and splicing destiny: on the cross-talk between chromatin structure and splicing.** *The EMBO journal* 2010, **29**:1629-1636.
7. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE: **High-resolution mapping and characterization of open chromatin across the genome.** *Cell* 2008, **132**:311-22.
8. Ling G, Sugathan A, Mazor T, Fraenkel E, Waxman DJ: **Unbiased, Genome-wide in vivo Mapping of Transcriptional Regulatory Elements Reveals Sex Differences in Chromatin Structure Associated with Sex-specific Liver Gene Expression.** *Molecular and cellular biology* 2010, MCB.00601-10-.
9. Micheli E, Martufi M, Cacchione S, Santis PDe, Savino M: **Self-organization of G-quadruplex structures in the hTERT core promoter stabilized by polyaminic side chain perylene derivatives.** *Biophysical chemistry* 2010.
10. Ioshikhes IP, Albert I, Zanton SJ, Pugh BF: **Nucleosome positions predicted through comparative genomics.** *Nature genetics* 2006, **38**:1210-5.
11. Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, LeProust EM, Hughes TR, Lieb JD, Widom J, Segal E: **The DNA-encoded nucleosome organization of a eukaryotic genome.** *Nature* 2009, **458**:362-6.
12. Chung H-R, Vingron M: **Sequence-dependent nucleosome positioning.** *Journal of molecular biology* 2009, **386**:1411-22.
13. Cui F, Zhurkin VB: **Structure-based analysis of DNA sequence patterns guiding nucleosome positioning in vitro.** *Journal of biomolecular structure & dynamics* 2010, **27**:821-41.
14. Luger K, Mäder AW, Richmond RK, Sargent DF, Richmond TJ: **Crystal structure of the nucleosome core particle at 2.8 A resolution.** *Nature* 1997, **389**:251-60.
15. Richmond TJ, Davey CA: **The structure of DNA in the nucleosome core.** *Nature* 2003, **423**:145-50.
16. Hall MA, Shundrovsky A, Bai L, Fulbright RM, Lis JT, Wang MD: **High-resolution dynamic mapping of histone-DNA interactions in a nucleosome.** *Nature structural & molecular biology* 2009, **16**:124-9.
17. Makde RD, England JR, Yennawar HP, Tan S: **Structure of RCC1 chromatin factor bound to the nucleosome core particle.** *Nature* 2010, **467**:562-566.
18. Zhang Y, Moqtaderi Z, Rattner BP, Euskirchen G, Snyder M, Kadonaga JT, Liu XS, Struhl K: **Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo.** *Nature structural & molecular biology* 2009, **16**:847-52.
19. Lowary PT, Widom J: **New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning.** *Journal of molecular biology* 1998, **276**:19-42.
20. Tolstorukov MY, Colasanti AV, McCandlish DM, Olson WK, Zhurkin VB: **A novel roll-and-slide mechanism of DNA folding in chromatin: implications for nucleosome positioning.** *Journal of molecular biology* 2007, **371**:725-38.
21. De Santis P, Morosetti S, Scipioni A: **Prediction of nucleosome positioning in genomes: limits and perspectives of physical and bioinformatic approaches.** *Journal of biomolecular structure & dynamics* 2010, **27**:747-64.
22. Trifonov EN, Sussman JL: **The pitch of chromatin DNA is reflected in its nucleotide sequence.** *Proceedings of the National Academy of Sciences of the United States of America* 1980, **77**:3816-20.
23. Xu F, Olson WK: **DNA architecture, deformability, and nucleosome positioning.** *Journal of biomolecular structure & dynamics* 2010, **27**:725-39.
24. Rohs R, West SM, Sosinsky A, Liu P, Mann RS, Honig B: **The role of DNA shape in protein-DNA recognition.** *Nature* 2009, **461**:1248-53.
25. Travers A, Hiriart E, Churcher M, Caserta M, Di Mauro E: **The DNA sequence-dependence of nucleosome positioning in vivo and in vitro.** *Journal of biomolecular structure & dynamics* 2010, **27**:713-24.

26. Arya G, Maitra A, Grigoryev SA: **A structural perspective on the where, how, why, and what of nucleosome positioning.** *Journal of biomolecular structure & dynamics* 2010, **27**:803-20.
27. Trifonov EN: **Nucleosome positioning by sequence, state of the art and apparent finale.** *Journal of biomolecular structure & dynamics* 2010, **27**:741-6.
28. Abeel T, Saeys Y, Bonnet E, Rouzé P, Van de Peer Y: **Generic eukaryotic core promoter prediction using structural features of DNA.** *Genome research* 2008, **18**:310-23.
29. Goñi JR, Fenollosa C, Pérez A, Torrents D, Orozco M: **DNAlive: a tool for the physical analysis of DNA at the genomic scale.** *Bioinformatics (Oxford, England)* 2008, **24**:1731-2.
30. Miele V, Vaillant C, D Aubenton-Carafa Y, Thermes C, Grange T: **DNA physical properties determine nucleosome occupancy from yeast to fly.** *Nucleic acids research* 2008, **36**:3746-56.
31. Goñi JR, Pérez A, Torrents D, Orozco M: **Determining promoter location based on DNA structure first-principles calculations.** *Genome biology* 2007, **8**:R263.
32. Lankas F, Sponer J, Langowski J, Cheatham TE: **DNA basepair step deformability inferred from molecular dynamics simulations.** *Biophysical journal* 2003, **85**:2872-83.
33. Morozov AV, Fortney K, Gaykalova DA, Studitsky VM, Widom J, Siggia ED: **Using DNA mechanics to predict in vitro nucleosome positions and formation energies.** *Nucleic acids research* 2009, **37**:4707-22.
34. Olson WK: **DNA sequence-dependent deformability deduced from protein-DNA crystal complexes.** *Proceedings of the National Academy of Sciences* 1998, **95**:11163-11168.
35. Araúzo-Bravo MJ, Fujii S, Kono H, Ahmad S, Sarai A: **Sequence-dependent conformational energy of DNA derived from molecular dynamics simulations: toward understanding the indirect readout mechanism in protein-DNA recognition.** *Journal of the American Chemical Society* 2005, **127**:16074-89.
36. Yamasaki S, Terada T, Shimizu K, Kono H, Sarai A: **A generalized conformational energy function of DNA derived from molecular dynamics simulations.** *Nucleic acids research* 2009, **37**:e135.
37. Fujii S, Kono H, Takenaka S, Go N, Sarai A: **Sequence-dependent DNA deformability studied using molecular dynamics simulations.** *Nucleic acids research* 2007, **35**:6063-74.
38. Flick JT, Eissenberg JC, Elgin SC: **Micrococcal nuclease as a DNA structural probe: its recognition sequences, their genomic distribution and correlation with DNA structure determinants.** *Journal of molecular biology* 1986, **190**:619-33.
39. Hörz W, Altenburger W: **Sequence specific cleavage of DNA by micrococcal nuclease.** *Nucleic acids research* 1981, **9**:2643-58.
40. Alexander M, Heppel LA, Hurwitz J: **The purification and properties of micrococcal nuclease.** *The Journal of biological chemistry* 1961, **236**:3014-9.
41. Pérez A, Lankas F, Luque FJ, Orozco M: **Towards a molecular dynamics consensus view of B-DNA flexibility.** *Nucleic acids research* 2008, **36**:2379-94.
42. Lavery R, Zakrzewska K, Beveridge D, Bishop TC, Case DA, Cheatham T, Dixit S, Jayaram B, Lankas F, Laughton C, Maddocks JH, 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 research* 2010, **38**:299-313.
43. Ozsolak F, Song JS, Liu XS, Fisher DE: **High-throughput mapping of the chromatin structure of human promoters.** *Nature biotechnology* 2007, **25**:244-8.
44. Field Y, Kaplan N, Fondufe-Mittendorf Y, Moore IK, Sharon E, Lubling Y, Widom J, Segal E: **Distinct modes of regulation by chromatin encoded through nucleosome positioning signals.** *PLoS computational biology* 2008, **4**:e1000216.
45. Bai L, Morozov AV: **Gene regulation by nucleosome positioning.** *Trends in genetics TIG* 2010, **26**:476-83.
46. Mavrich TN, Ioshikhes IP, Venters BJ, Jiang C, Tomsho LP, Qi J, Schuster SC, Albert I, Pugh BF: **A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome.** *Genome research* 2008, **18**:1073-83.
47. Segal E, Fondufe-Mittendorf Y, Chen L, Thåström A, Field Y, Moore IK, Wang J-PZ, Widom J: **A genomic code for nucleosome positioning.** *Nature* 2006, **442**:772-8.
48. Lee W, Tillo D, Bray N, Morse RH, Davis RW, Hughes TR, Nislow C: **A high-resolution atlas of nucleosome occupancy in yeast.** *Nature genetics* 2007, **39**:1235-44.
49. Albert I, Mavrich TN, Tomsho LP, Qi J, Zanton SJ, Schuster SC, Pugh BF: **Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome.** *Nature* 2007, **446**:572-6.
50. Yuan G-C, Liu Y-J, Dion MF, Slack MD, Wu LF, Altschuler SJ, Rando OJ: **Genome-scale identification of nucleosome positions in S. cerevisiae.** *Science (New York, N.Y.)* 2005, **309**:626-30.
51. Yuan GC, Liu YJ, Dion MF, Slack MD, Wu LF, Altschuler SJ, Rando OJ: **Genome-scale identification of nucleosome positions in S. cerevisiae.** *Science (New York, N.Y.)* 2005, **309**:626-3010, 1126/science.1112178.
52. Lee W, Tillo D, Bray N, Morse RH, Davis RW, Hughes TR, Nislow C: **A high-resolution atlas of nucleosome occupancy in yeast.** *Nature genetics* 2007, **39, 1235-4410.1038/ng2117.**
53. Field Y, Kaplan N, Fondufe-Mittendorf Y, Moore IK, Sharon E, Lubling Y, Widom J, Segal E: **Distinct modes of regulation by chromatin encoded through nucleosome positioning signals.** *PLoS computational biology* 2008, **4**:e100021610, 1371/journal.pcbi.1000216.
54. Albert I, Mavrich TN, Tomsho LP, Qi J, Zanton SJ, Schuster SC, Pugh BF: **Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome.** *Nature* 2007, **446**:572-610.1038/nature05632.
55. Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, LeProust EM, Hughes TR, Lieb JD, Widom J, Segal E: **The DNA-encoded nucleosome organization of a eukaryotic genome.** *Nature* 2009, **458**:362-610, 1038/nature07667.
56. Chung H-R, Dunkel I, Heise F, Linke C, Krobitsch S, Ehrenhofer-Murray AE, Sperling SR, Vingron M: **The effect of micrococcal nuclease digestion on nucleosome positioning data.** *PloS one* 2010, **5**:e1575410, 1371/journal.pone.0015754.
57. Rando OJ, Ahmad K: **Rules and regulation in the primary structure of chromatin.** *Current opinion in cell biology* 2007, **19**:250-6.
58. Rohs R, Jin X, West SM, Joshi R, Honig B, Mann RS: **Origins of specificity in protein-DNA recognition.** *Annual review of biochemistry* 2010, **79**:233-69.
59. Peckham HE, Thurman RE, Fu Y, Stamatoyannopoulos JA, Noble WS, Struhl K, Weng Z: **Nucleosome positioning signals in genomic DNA.** *Genome research* 2007, **17**:1170-7.
60. Weiner A, Hughes A, Yassour M, Rando OJ, Friedman N: **High-resolution nucleosome mapping reveals transcription-dependent promoter packaging.** *Genome research* 2010, **20**:90-100.
61. Locke G, Tolkunov D, Moqtaderi Z, Struhl K, Morozov AV: **High-throughput sequencing reveals a simple model of nucleosome energetics.** *Proceedings of the National Academy of Sciences* 2010, 1003838107.
62. Segal E, Widom J: **From DNA sequence to transcriptional behaviour: a quantitative approach.** *Nature reviews Genetics* 2009, **10**:443-56.
63. Fan X, Moqtaderi Z, Jin Y, Zhang Y, Liu XS, Struhl K: **Nucleosome depletion at yeast terminators is not intrinsic and can occur by a transcriptional mechanism linked to 3'-end formation.** *Proceedings of the National Academy of Sciences of the United States of America* 2010, **107**:17945-50.
64. Zhang Z, Wippo CJ, Wal M, Ward E, Korber P, Pugh BF: **A Packing Mechanism for Nucleosome Organization Reconstituted Across a Eukaryotic Genome.** *Science* 2011, **332**:977-98010, 1126/science.1200508.
65. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nature genetics* 2000, **25**:25-9.
66. Mavrich TN, Jiang C, Ioshikhes IP, Li X, Venters BJ, Zanton SJ, Tomsho LP, Qi J, Glaser RL, Schuster SC, Gilmour DS, Albert I, Pugh BF: **Nucleosome organization in the Drosophila genome.** *Nature* 2008, **453**:358-62.
67. Mavrich TN, Jiang C, Ioshikhes IP, Li X, Venters BJ, Zanton SJ, Tomsho LP, Qi J, Glaser RL, Schuster SC, Gilmour DS, Albert I, Pugh BF: **Nucleosome organization in the Drosophila genome.** *Nature* 2008, **453**:358-6210, 1038/nature06929.
68. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome biology* 2009, **10**:R25.
69. Flores O, Orozco M: **nucleR: a package for non-parametric nucleosome positioning.** *Bioinformatics (Oxford, England)* 2011, **27**:2149-215010, 1093/bioinformatics/btr345.

70. Pérez A, Marchán I, Svozil D, Sponer J, Cheatham TE, Laughton CA, Orozco M: **Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers.** *Biophysical journal* 2007, **92**:3817-29.

doi:10.1186/1471-2164-12-489
Cite this article as: Deniz *et al.*: Physical properties of naked DNA influence nucleosome positioning and correlate with transcription start and termination sites in yeast. *BMC Genomics* 2011 **12**:489.

1.2. Impact of Methylation on the Physical Properties of DNA

Previously, we introduced the role of the epigenetic modification of the cytosine caused by methylation in mammalian genomes (page 10). In this work, we created an additional set of basepair parameters to the already available ones (Lankas *et al.*, 2003; Pérez, Marchán, *et al.*, 2007; Lavery *et al.*, 2010b), which describe the physical properties of basepairs in a methylated CpG context.

These new MD-derived physical parameters predicted an increased stiffness of the CpG steps upon methylation. This theoretical conclusion was then experimentally validated by a circularization assay. Double stranded normal and methylated DNA short-fragments were ligated and the ratio between linear and circular DNA was calculated. Methylated sample showed a lower ratio of circularization, consistent with an increased stiffness as predicted by the computational approach.

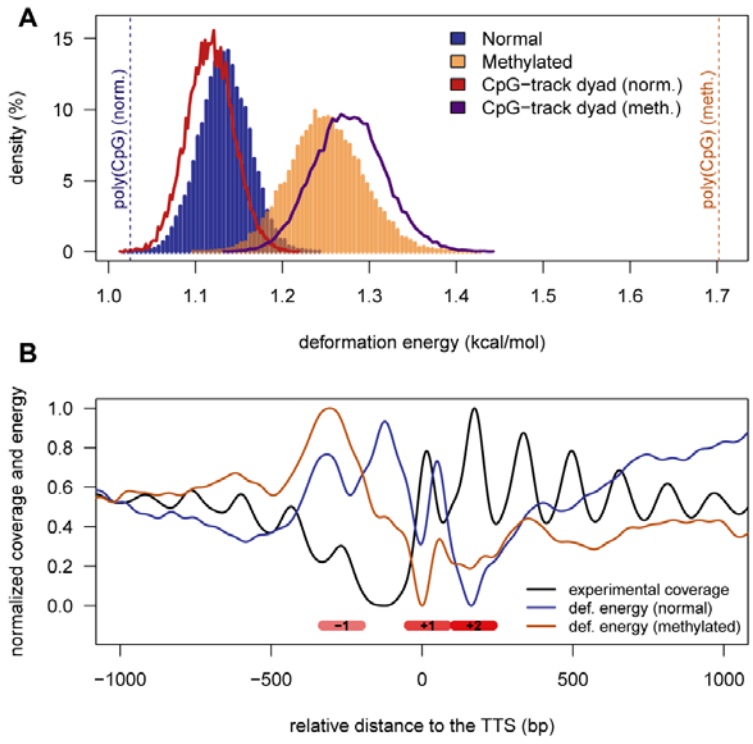


Fig. 19. Theoretical effects of methylated DNA in nucleosome formation. A) Deformation energy distribution for normal and methylated DNA in a random set of genomic sequences. Also special sequences with a CpG on the nucleosome dyad (in base 74) and a poly(CpG) sequence are shown as references. B) Experimental nucleosome coverage (black) is compared with the averaged deformation energy, with normal (blue) and methylated (orange) sequences.

Furthermore, the integration of these new parameters to the deformation energy model presented previously (page 33) predicts a change in the nucleosome positioning on the gene promoters accordingly to the methylation state of the DNA (Fig. 19). Despite the drop in the affinity to form nucleosomes after DNA methylation was confirmed by a reconstitution assay, the real effects *in vivo* of these theoretical changes would require an experimental validation on mammal cells.

Publication:

Pérez, A., Castellazzi, C. L., Battistini, F., Collinet, K., Flores, O., Deniz, Ö., Ruiz, M. L., Torrents, D., Eritja, R., Soler-López, M., Orozco, M. (2012) "*Impact of Methylation on the Physical Properties of DNA*", Biophysical Journal 102: 2140–8.

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

Impact of Methylation on the Physical Properties of DNA

Alberto Pérez,[†‡△] Chiara Lara Castellazzi,[‡△] Federica Battistini,[‡] Kathryn Collinet,[‡] Oscar Flores,[‡] Ozgen Deniz,[‡] Maria Luz Ruiz,[‡] David Torrents,[†¶] Ramon Eritja,[§] Montserrat Soler-López,[‡] and Modesto Orozco[†‡‖*]

[†]IRB-BSC Joint Research Program on Computational Biology, Barcelona Supercomputing Center, Barcelona, Spain; [‡]IRB-BSC Joint Research Program on Computational Biology and [§]Chemistry and Molecular Pharmacology Program, Institute for Research in Biomedicine, Barcelona, Spain; [¶]Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain; and [‖]Departamento de Bioquímica, Facultat de Biología, Barcelona, Spain

ABSTRACT There is increasing evidence for the presence of an alternative code imprinted in the genome that might contribute to gene expression regulation through an indirect reading mechanism. In mammals, components of this coarse-grained regulatory mechanism include chromatin structure and epigenetic signatures, where d(CpG) nucleotide steps are key players. We report a comprehensive experimental and theoretical study of d(CpG) steps that provides a detailed description of their physical characteristics and the impact of cytosine methylation on these properties. We observed that methylation changes the physical properties of d(CpG) steps, having a dramatic effect on enriched CpG segments, such as CpG islands. We demonstrate that methylation reduces the affinity of DNA to assemble into nucleosomes, and can affect nucleosome positioning around transcription start sites. Overall, our results suggest a mechanism by which the basic physical properties of the DNA fiber can explain parts of the cellular epigenetic regulatory mechanisms.

INTRODUCTION

Determining the mechanisms that regulate gene expression in complex organisms is the next frontier of genomics research (1). In the traditional paradigm, specific proteins regulate gene expression through the recognition of certain sequence signals (by means of specific hydrogen-bond interactions) upstream of the transcription start sites (TSSs) (2). Nevertheless, there is increasing evidence about the presence of an alternative code that may contribute to a rough regulation of gene expression through an indirect reading mechanism, probably signaled by chromatin structure and epigenetic marks (3,4). This mechanism is unlikely (even in a synergistic manner) to achieve the fine-tuning and specificity of the direct protein-DNA readout. Conversely, it probably plays a pivotal role in basal gene expression, which requires less regulation and for which the extreme cost of developing a highly specific protein regulation infrastructure seems unjustified. Key players in this regulatory mechanism may be d(CpG) steps, which despite being largely underrepresented in the genome of complex organisms are enriched in nearly 60% of human promoters, where they often define ultrarich CpG segments, the so-called CpG islands (5,6). Even if CpG islands do not contain specific transcription binding motifs, they clearly favor the downstream binding of the transcription machinery (7), particularly for those genes that are usually active. The molecular basis of the d(CpG) effect on gene regulation remains unclear, although it has been suggested to be related to the definition of DNA fiber properties (8).

One of the most intriguing features of the d(CpG) step is its ability to undergo nonmutagenic chemical modifications such as cytosine methylation (9). In mammalian genomes, DNA methyltransferases (DNMTs) can transfer a methyl group from S-adenosylmethionine to cytosine at CpG dinucleotides (10). The bulk of the methylation takes place during DNA replication in the S-phase of the cell cycle (11), and is the most abundant form of post-replicative DNA modification observed in eukaryotic organisms (12). During this process, the cytosine is flipped 180° out of the DNA backbone into an active-site pocket of the enzyme (13) where methylation of cytosine takes place.

Intriguingly, methylation of cytosines seems to be an erroneous decision of evolution because it dramatically increases the chances of C→T mutation, but this seeming disadvantage is compensated for by the gain in regulatory possibilities offered by methylation. Indeed, highly methylated DNA is typically associated with inactive genes, whereas methylation depletion is observed for active genes (14,15). Furthermore, most cytosines in CpG steps, except those in CpG islands, are methylated in vertebrate somatic cells (16,17). The first step of methylation occurs early in mammalian development as a result of de novo DNMTs (Dnmt3a and Dnmt3b) (18) that methylate CpG steps in both DNA strands. The methylation profile is conserved by maintenance DNA methyltransferase (Dnmt1) throughout cell divisions. During replication, daughter strands are nonmethylated, resulting in hemimethylated DNA. Dnmt1 recognizes hemimethylated CpG steps and methylates the daughter strand (19). Recent studies have demonstrated that changes in methylation patterns along CpG islands and CpG shores (methylation

Submitted December 27, 2011, and accepted for publication March 22, 2012.

△Alberto Pérez and Chiara Lara Castellazzi contributed equally to this work.

*Correspondence: modesto@mmb.pcb.ub.es

Editor: Michael Levitt.

0006-3495/12/05/2140/9 $2.00

doi: 10.1016/j.bpj.2012.03.056

hotspots on the outskirts of CpG islands (20)) correlate with tissue differentiation and cancer, proving the role of methylation in cellular reprogramming (20,21,23,24). The regulatory function of methylated cytosines (hereafter referred to as ^{Me}C) was traditionally explained by their interaction with methyl-CpG binding domain proteins (MBDs) (18), but considering the prevalence of cytosine methylation, MBDs alone cannot entirely account for the role of d(CpG) methylation in the cell. An increased level of complexity is provided by the almost nonexistent sequence specificity of the DNMTs, precluding the mechanism underlying the methylation reaction (26).

Here we present a comprehensive theoretical analysis of the d(CpG) properties along with an experimental validation of key theoretical findings. We show that d(CpG) steps display unique physical properties, especially in the context of CpG islands, that severely change upon methylation. Our calculations suggested, and experiments confirmed, that DNA segments containing ^{Me}C are very stiff and hard to bend, and display a lower tendency to circularize or form nucleosomes by wrapping around histones. The latter effect has striking consequences for the organization of nucleosome arrays near TSSs, which in turn modifies the accessibility of regulatory proteins, leading to alterations in the pattern of DNA expression. Overall, without diminishing the role of specific regulatory proteins, basic descriptors of DNA physical properties can help us rationalize several seemingly disconnected pieces of the puzzle of DNA regulation.

METHODS

Molecular-dynamics simulations

We performed molecular-dynamics (MD) simulations on an array of different oligomers containing CpG steps, in both their methylated and nonmethylated forms (see Table S1 in the Supporting Material). We also included in our analysis trajectories of unmethylated CpG step data from the Ascona B-DNA Consortium (ABC) (27) database in all tetramer environments to enrich the dynamics database. All simulations were carried out in duplicates for 100 ns (after equilibration) using explicit solvent, the parmbsc0 refinement of the Amber force field (28), and state-of-the-art simulation conditions (Supporting Material).

Mesoscopic model of DNA flexibility

We derived a flexibility model from different MD equilibrium trajectories using a harmonic model (29–31). Accordingly, we projected the MD trajectories onto a helical reference system to obtain equilibrium values and derive the covariance matrix, which we then inverted to recover the stiffness matrices for each basepair step, from which a mesoscopic estimate of the energy associated to a given deformation can be easily computed (29,32) as

$$E = 0.5\sum_{i=1}^{6}\sum_{j=1}^{6} f_{ij}\Delta X_i \Delta X_j,$$

where ΔX_i is the perturbation from equilibrium geometry (Fig. 1), and f_{ij} are elements of Θ, where Θ is the 6×6 stiffness matrix expressing the stiffness of a given step to deformation in roll, tilt, twist, slide, shift, and rise (see Supporting Material, Fig. 2, and Pérez et al. (33)). Alternatively, global deformation parameters can be derived using a similar approach, but considering global instead of local geometric descriptors (see Fig. S1, Supporting Material, and Lankas et al. (34)). Note that elastic parameters derived from protein-DNA crystal complexes (29,31) or simulation data are in good agreement (33), supporting their use to describe DNA flexibility. Here, we favored the use of MD-derived values for consistency with the newly derived parameters describing ^{Me}CpG steps (which cannot be derived from analysis of crystal structures).

In this work, we used a mesoscopic method to estimate deformation energy related to nucleosome formation and circularization assays. This implies that indirect readout mechanisms prevail over the direct readout for the description of sequence preferences in nucleosome binding. A second implication is that these deformations follow a harmonic behavior. Both of these assumptions represent simplifications, and thus the validity of the method is not always guaranteed (32,35,36).

Circularization assays and modeling

We carried out DNA circularization experiments to validate our theoretical estimations about the impact of methylated cytosines on DNA physical properties. For this purpose, we designed a short polymerizable oligonucleotide (d(GAAAAAACGGGCGAAAAACGG)·d(TCCCGTTTTTCGCCCGTTTTT)) based on a reported sequence favoring the formation of minicircles (37), with the incorporation of a central CpG dinucleotide subject to be methylated and 5′-sticky ends to enable the formation of multimers. Thus, under favorable ligation conditions, the multimers form circles that are as short as allowed by the geometry and flexibility of the DNA (Fig. 3). As a negative control, we selected a previously reported nonbendable oligonucleotide (d(GCAAATATTGAAAAC)·d(GCGTTTTCAATATTT; see Supporting Material for details). The ligation products were analyzed by atomic force microscopy (AFM) and by two-dimensional gel electrophoresis (Fig. 3, Fig. S2, Fig. S3, and Supporting Material).

Circularization efficiency was determined based on the *J*-factor, which defines the ratio of circular and linear DNA species for a given sequence length (37). Experimental *J*-factors were extracted from the linear and circle DNA signal intensities as detected on two-dimensional gels (see Supporting Material and Fig. S3). Although the absolute *J*-factors depend on the experimental setup, the ratio of *J*-factors of methylated versus nonmethylated oligos provide a reliable measure of the impact of methylation on DNA circularization. Accordingly, experimental *J*-factors can validate whether theoretical suggestions regarding physical changes induced by methylation are correct. We derived theoretical *J*-factors from Monte Carlo simulations (Supporting Material) using mesoscopic descriptors derived from MD simulations (see above).

Mesoscopic model of nucleosome deformation energy

We theoretically determined the ability of a 147-mer DNA sequence to wrap around a nucleosome using harmonic deformation energy as described above considering mesoscopic descriptors. To reduce the noise, we determined the deformation vector (ΔX above) using the target geometry obtained by Fourier-averaging all available crystal structures of nucleosome particles. As noted above, our mesoscopic model is useful in so far as histone-bound DNA deforms harmonically and the indirect readout has an important contribution in directing nucleosome formation. Regarding the first point, the rotational degrees of motion in nucleosomes (twist, roll, and tilt) clearly fall within the normal fluctuations of DNA (38,39), and only the translational parameter slide shows slightly more positive values than expected. Clearly, evolution has optimized nucleosome positioning sequences to have flexible steps such as d(CpA) and d(TpA) in crucial positions to accommodate the deformations required for nucleosome binding (see results of SELEX experiments in Thåström et al.

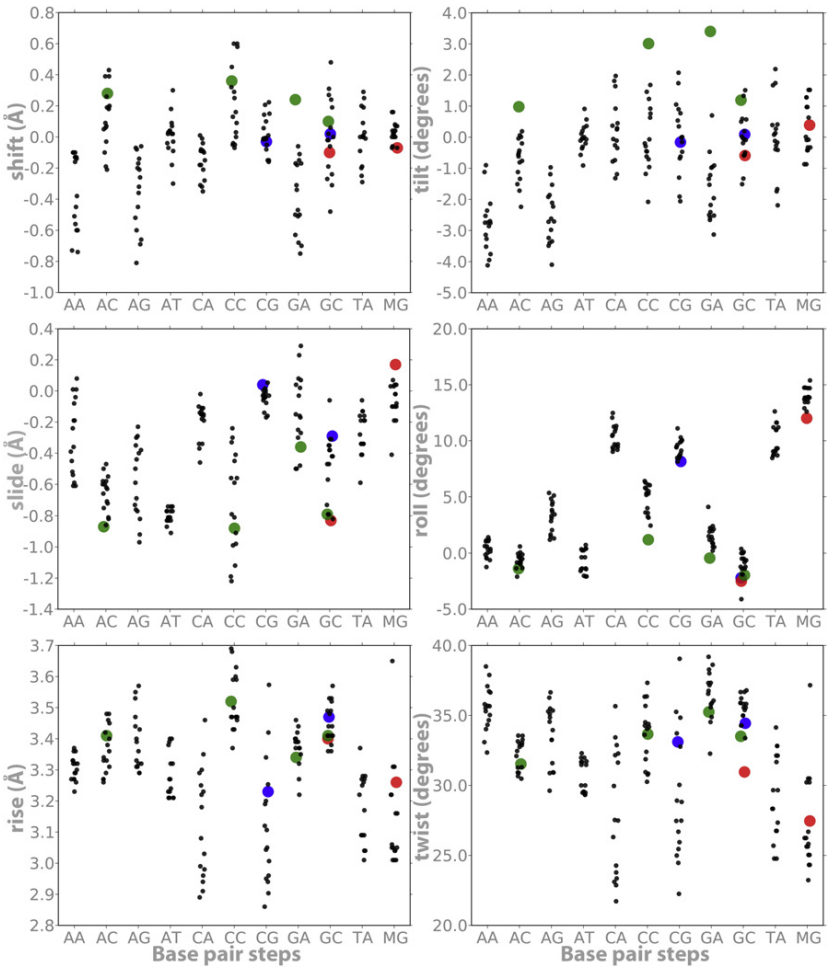


FIGURE 1 Average helical parameters (translations in angstroms and rotations in degrees) derived from MD simulations of the usual 10 dinucleotides plus d(^{Me}CpG; referred to as MG in the figure). The black dots correspond to control simulations (those performed for this work and those obtained from the ABC database (27)) for each central basepair step representing the different tetramer environments; notice the slight displacement from the vertical to avoid stacking of data. Blue dots stand for d(CpG) and d(GpC) steps when embedded in a poly d(CpG) track. Green dots refer to neighboring methylated steps (Xp^{Me}C): d(Ap^{Me}C), d(Tp^{Me}C)=d(GpA), d(Cp^{Me}C), and d(Gp^{Me}C). Finally, red dots stand for d(^{Me}CpG) and d(Gp^{Me}C) in the context of a poly d(CpG) track.

(40)). Note also that by using a single conformation to model all nucleosomes, Tolstorukov et al. (38) and Balasubramanian et al. (39) found a high degree of correlation between the predictions arising from harmonic deformation energies and nucleosome positioning sequences, a result that was consistent with nucleosome location experiments performed by our group (41). Thus, without ignoring its limitations, we believe the model can be useful to rationalize nucleosome positioning.

In vitro nucleosome reconstitution

To assess the effect of methylated DNA on nucleosome assembly, we selected a nucleosome positioning sequence (DNA construct 601.2 in Anderson and Widom (42)) to reconstitute nucleosomes in vitro after incubation with histones, before and after extensive DNA methylation (see Supporting Material). Methylated states were verified by DNA sequencing. The reconstituted nucleosomes were subsequently analyzed by gel shift assays (see Fig. 4).

RESULTS

Physical properties of CpG steps and CpG islands

MD simulations performed here in conjunction with those retrieved from the ABC database (27) revealed that sequence is crucial for defining the DNA equilibrium geometries (Fig. 1). In general, Pyr-Pur steps show a lower rise and twist, as well as higher roll values, than the rest of the dinucleotide steps. Additionally, they display an unusually large dispersion in certain key equilibrium helical parameters (e.g., twist), arising from different tetramer environments. Focusing on the different tetramer environments, we can see that the d(CpG) steps are peculiar in presenting bimodal distributions of some parameters (e.g., twist; see Fig. S4), which confirms previous ABC findings (27). Such bimodality is not an artifact that arises from incomplete sampling, because it is also present in multi-microsecond trajectories (43), and suggests that the d(CpG) step is specially flexible, as confirmed by a stiffness analysis (Fig. 2). It is worth noting that the large deformability in twist and roll, combined with large roll values, suggests that protein-induced curvature may be favored in DNA with CpG steps. Fig. S5 shows that CpG, despite the neighboring basepair, is more curved than most basepair steps (only TA and CA are comparable) and is directed preferentially toward the major groove.

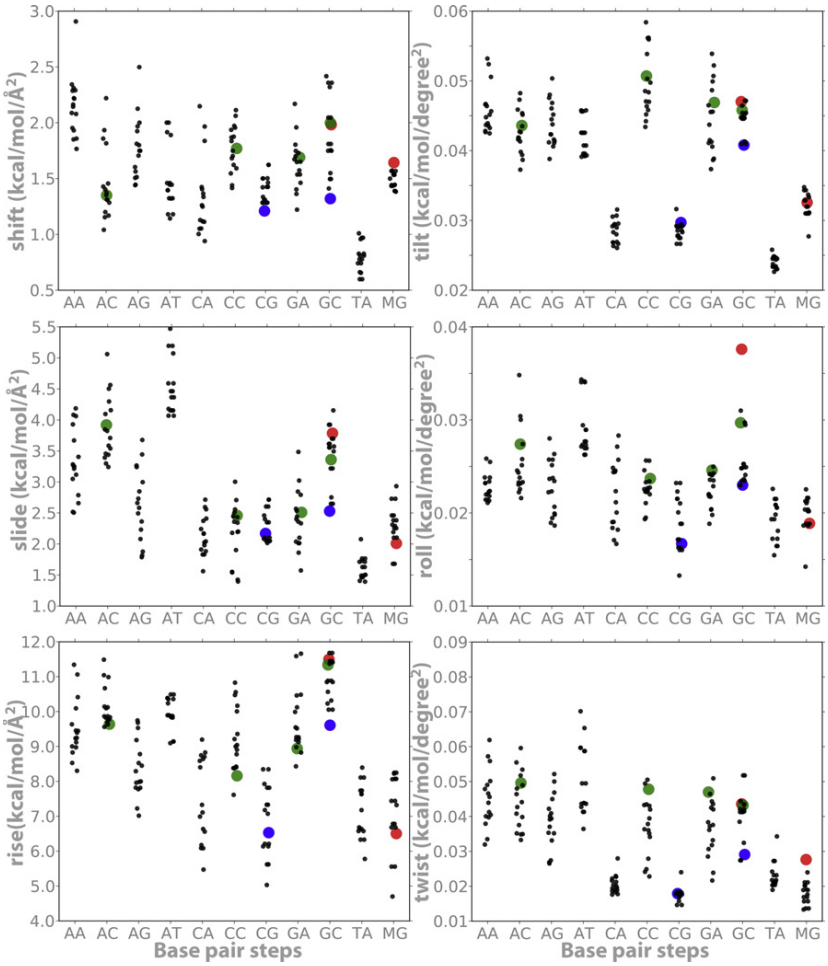


FIGURE 2 Average stiffness helical parameters (translations in kcal/mol Å^2 and rotations in kcal/mol deg^2) derived from MD simulations of the usual 10 dinucleotides plus d(^{Me}CpG; referred to as MG in the figure). Same notation as in Fig. 1.

One might expect the unique physical properties of d(CpG) steps to be amplified in long d(CpG)$_n$ tracks (i.e., CpG islands). Surprisingly, this is not the case, and the global geometrical properties of long d(CpG) segments are different from those expected by extrapolating the individual characteristics of the d(CpG) steps (Fig. S1). Thus, the high flexibility of the d(CpG) steps suggests that poly d(CpG) should be extremely flexible. Conversely, the d(CpG)$_9$ segment studied here is hardly distinguishable from other 18-mer duplexes in terms of global unwinding and isotropic bending. This is hardly surprising when one considers the differences between an individual d(CpG) step and a poly d(CpG). The former has the properties of an individual d(CpG) step, whereas the d(CpG)$_9$ segment has properties due to the alternation of d(CpG) and d(GpC) steps. Thus, the lower flexibility of d(GpC) steps (Fig. S1) makes the whole sequence overall stiffer than a simple extrapolation of d(CpG) properties. Furthermore, whereas the large roll of individual d(CpG) steps would imply a strong curvature of the entire oligonucleotide, the d(CpG)$_9$ curvature is actually very moderate due to the low roll values of d(GpC) steps. In conclusion, the properties of long d(CpG) segments are distinct from the extrapolation of properties of isolated d(CpG) steps, warning against the use of oversimplified rules of DNA flexibility.

Effect of CpG methylation

Early structural experiments suggested that cytosine methylation (Fig. S5) might induce helical transitions from B- to Z-DNA (44). However, a secondary structure analysis of CpG methylated oligonucleotides by circular dichroism spectroscopy (Supporting Material and Fig. S6) revealed that the transition only occurred at nonphysiological salt concentrations (from 1 to 2 M NaCl). This evidence is in agreement with our MD results and previously reported Fourier transform infrared (FTIR) spectroscopy data (45), demonstrating that in vivo DNA remains in the B-form upon methylation, and accordingly, the transition to the Z-form is not the underlying determinant for the physiological role of CpG methylation.

The results of the MD simulations suggest that when methylated, d(CpG) steps increase their average roll value and reduce their twist (Fig. 1), leading to an increase in local

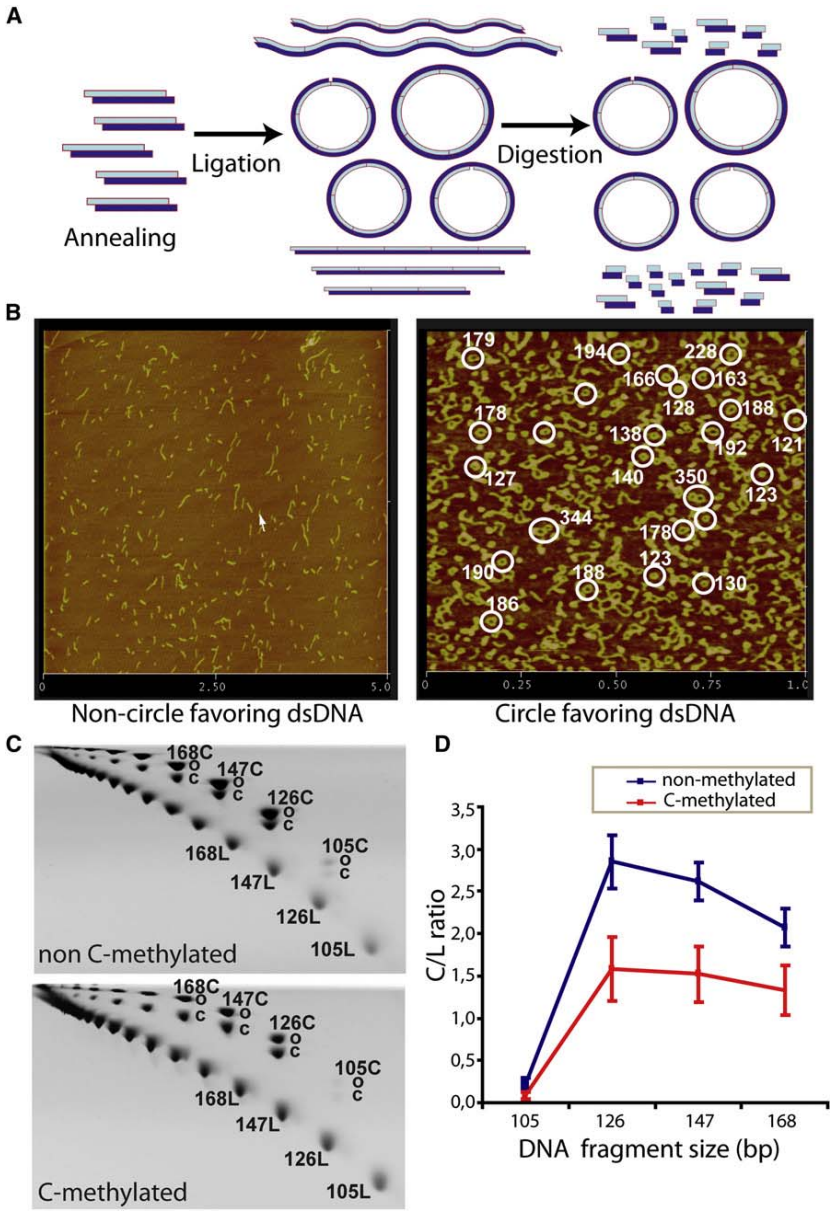


FIGURE 3 Overview of circularization assays. (*A*) Schematic diagram of the underlying principle of circularization assays. A DNA oligonucleotide is first annealed to form duplexes and subsequently is multimerized-circularized by a ligation reaction. Under favorable ligation conditions, DNA forms circles as short as allowed by the geometry and flexibility of the DNA. Only circularized DNA is resistant to an exonuclease digestion; linear multimers will be degraded. (*B*) AFM images of ligation products for 15-bp nonfavoring (*left*) and 21-bp favoring (*right*) circularization oligonucleotides; circle size estimations are highlighted in white. (*C*) 2D polyacrylamide native gels showing different migrations of linear (L) and circular (C) DNA species (which can be either covalently closed (cC) or nicked open (oC)) for nonmethylated and ^{Me}C oligomers of 21 bp, respectively. Linear DNA molecules are positioned on the lower diagonal, and circular DNA molecules are positioned on the upper diagonal. (*D*) The circularization efficiency is expressed as the ratio between C and L molecules of the same size.

curvature. Furthermore, methylation makes d(CpG) steps stiffer, especially in terms of roll and tilt deformations: the ^{Me}CpG step has larger tilt and roll force constants on average than the CpG step (MG and CG, respectively, in Fig. 2). Methylation also alters the geometric properties of the basepair step previous to the d(CpG) site, here denoted as d(XpC), where X = A, C, G, or T (Figs. 1 and 2). In canonical DNA, d(XpC) steps tend to compensate for the geometry and relative stiffness of d(CpG) in twist, tilt, and roll. However, upon methylation, we observed an increase of force constants for rotational parameters in both the ^{Me}CpG and d(Xp^{Me}C) steps (*green dots* in Fig. 2). Hence, the additive effect of methylation leads to significant alterations in the global physical properties of DNA, especially for CpG islands (Figs. 1 and 2, and Fig. S1).

The higher stiffness of the d(^{Me}CpG) steps should lead to a decrease in the DNA circularization efficiency, which must be especially visible for the smallest circles. Indeed, Monte Carlo calculations using the MD-derived stiffness parameters (see Materials and Methods, and Supporting Material) suggested that circularization of 126- to 189-bp-long methylated oligos (one ^{Me}CpG every 21 bp) is more difficult than circularization of unmethylated ones (with a relative *J*-factor of 0.05–0.2). This result was confirmed by circularization experiments on the same sequence with a relative *J*-factor of ~0.5 (see Fig. 3, Materials and Methods, and Supporting Material).

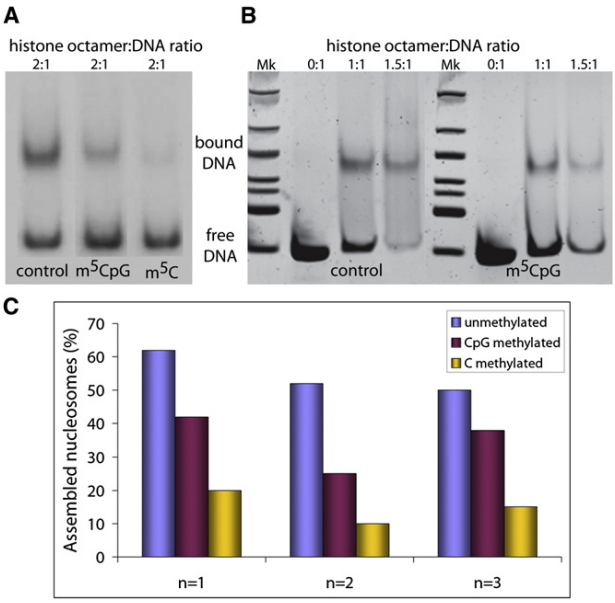


FIGURE 4 In vitro nucleosome core particle reconstitution. Results of gel mobility shift assays of nucleosomes reconstituted in vitro with a 147-bp 601.2 DNA fragment containing either C or ^{Me}C at different histone octamer concentrations are shown. The upper bands correspond to histone core-bound DNA, and lower bands correspond to unbound (or free) DNA. Mk: DNA ladder for size band estimation. (*A*) Radiolabeled DNA bands. (*B*) DNA bands stained with SyBr Safe (Invitrogen) and visualized by ultraviolet light. (*C*) Histograms displaying the percentage of in vitro assembled nucleosomes using the same sequence in different methylation conditions coming from triplicate experiments.

Potential impact on chromatin structure

The sequence-dependent physical properties of DNA play a crucial role in determining nucleosome positioning (46–52) and thus are instrumental in genome regulation (53,54). An analysis of nucleosome distribution in *Saccharomyces cerevisiae* (55,56) revealed that d(CpG)s are enriched in nucleosome-bound regions but depleted in internucleosomal segments, supporting the idea that d(CpG) steps are easily fitted into the nucleosome structure. Protein-DNA interactions are governed by direct and indirect readouts. The former arises from protein-DNA direct interactions, whereas the latter is based on the ability of a DNA sequence to deform into a conformation that makes the interaction happen. Without underestimating the importance of direct readout mechanisms in nucleosome binding, we note that indirect readout models seem to capture well the global positioning profile of nucleosomes (41). Thus, changes in the physical properties of the DNA fiber related to methylation should have a direct impact on nucleosome affinity and positioning. Our models prompted us to hypothesize that in the absence of external factors (e.g., MBD proteins or chromatin remodelers), the increased stiffness due to d(CpG) methylation leads to a higher deformation energy required to wrap DNA around a nucleosome (Fig. 5). We tested this hypothesis by conducting in vitro nucleosome reconstitution experiments (see Materials and Methods, and Supporting Material) with normal and methylated DNAs. The results confirm that the d(^{Me}CpG) DNA has a lower ability to form nucleosomes than the nonmethylated sequence (Fig. 4).

Interestingly, nucleosome formation was further decreased when all cytosines in the DNA were methylated, which confirms that a reduced flexibility is mainly responsible for the lower affinity of methylated DNA for the histones. Other nonmammalian organisms that have alternative cytosine methylation patterns besides the methylation of CpG steps could use this strategy for gene regulation. Tillo and Hughes (57) established that increasing C+G contents correlate well with higher nucleosome formation. Therefore, it is not surprising that a mechanism in which more cytosines are methylated would further rigidify the sequence, further changing the nucleosome positioning preferences (57).

DISCUSSION

d(CpG) steps are statistically underrepresented in the genome, but they appear concentrated in regulatory regions,

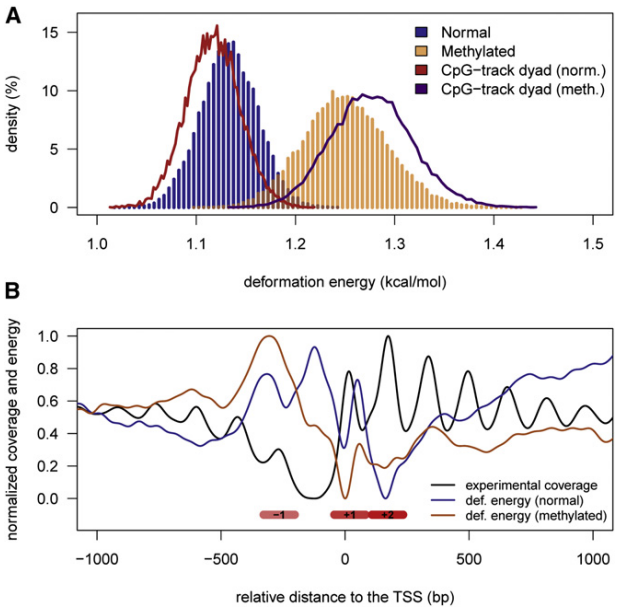


FIGURE 5 Impact of methylation on nucleosome positioning. (*A*) The distribution of predicted energies (per base step) for 147-bp-long random DNAs in normal (*blue histogram*) and methylated (*orange histogram*) forms, respectively. The curves correspond to the predicted energy (per base step) when random oligos contain a poly d(CpG) track at the dyad, in normal (*red*) and methylated (*magenta*) forms. (*B*) The nucleosome distribution surrounding the TSSs of yeast genes determined from MNase digestion experiments (*black line*), compared with the predicted distortion energy to wrap a nucleosome in those sites when genomic DNA is normal (*blue line*) or methylated (*orange line*). All values were normalized to facilitate the interpretation of the plots. Nucleosome positions −1, +1, and +2 are indicated by red boxes for clarity.

which suggests that d(CpG) steps provide the suitable physical properties that enable the DNA to efficiently interact with regulatory proteins (58,59) and help define the correct nucleosome positioning. Indeed, our MD results suggest that isolated d(CpG) steps are curved (see Fig. S5) and particularly flexible, making d(CpG) steps appropriate for those regions in which DNA needs to be locally distorted to facilitate protein binding, in agreement with early NMR measures (60). This hypothesis is supported by an analysis of whole-genome nucleosome positioning in yeast, which revealed a d(CpG) step depletion in internucleosomal segments (41,55,56).

Mesoscopic calculations (Fig. 5 *A*) indicate that a $d(CpG)_5$ segment located at the dyad axis may favor nucleosome formation, and hence very long d(CpG) tracks (CpG islands) are very likely to assemble into nucleosomes. Nucleosomes can easily move along the d(CpG) track, presumably leading to a nucleosome-depleted region at the external borders of the island, and thereby imprinting a distinct nucleosome array organization that would define the accessibility to regulatory regions downstream of the CpG islands, where many promoters are located. This also accords with the fact that high C+G content correlates with nucleosome positioning (57)

Methylation increases the curvature of d(CpG) steps (Fig. S5), although this local geometric effect tends to be compensated for by neighboring steps. In fact, all tested 18-mer methylated oligos (except CpG islands) were less curved and less flexible than their unmethylated counterparts. On the other hand, as previously suggested (61,62), methylation increases d(CpG) stiffness, and this effect propagates to neighboring steps, leading to a global increase in the rigidity of DNA. Our results suggest that this effect alone explains (within the indirect readout model) the limited ability of methylated DNA to interact with certain proteins, such as transcription factors (63–65). Our in silico simulations and in vitro nucleosome reconstitution experiments showed that methylation reduces DNA affinity for nucleosomes, probably due to the increased rigidity of the DNA fiber. Our results are in full agreement with previous findings (66,67), with recent data about the anticorrelation between nucleosome formation and methylation (68), and with physical intuition that suggests that a more flexible fiber should wrap more easily than a rigid one. Additionally, the rigidifying effect of ^{Me}CpG is also observed in recent fluorescence resonance energy transfer (FRET)-derived data (69,70) showing that methylating nucleosome-bound DNA results in nucleosome compaction and rigidity. Taken together, these results indicate that DNA methylation may regulate nucleosome dynamics by increasing the rigidity of DNA either before or during nucleosome assembly.

Our combined theoretical and experimental results demonstrate that methylation decreases nucleosome formation. Nucleosome depletion is usually considered as a signal of gene activity (71,72). At the same time, gene activity correlates with low levels of methylation (73). We postulate that the presence of almost 10^5 ^{Me}CpG steps present in the genome could significantly modify the nucleosome positioning landscape. This hypothesis would explain how the gene expression pattern can change while the number of nucleosomes (but not positions) is kept constant in either methylated or canonical genomes. Hence, we analyzed the nucleosome organization around TSSs on the unmethylated yeast genome and subsequently compared the in silico effects of methylation on nucleosome positioning (Fig. 5 *B*). TSSs are typically characterized by a nucleosome-depleted region and well-positioned −1, +1, and +2 nucleosomes (55,74). As expected, these positions are clearly marked in the energy profiles: regions with high deformation energy signal nucleosome-depleted areas and vice versa (Fig. 5). These profiles support recent claims about particular nucleosome positioning sites (signaled by large in vitro propensities for nucleosome assembly) anchoring the formation of nucleosomal arrays in vivo (75,76). Methylation of d(CpG) steps modifies the deformation energy profile associated with nucleosome wrapping around the TSS, which ultimately may be reflected by a change in the nucleosome array (Fig. 5 *B*; note that this does not necessarily make the nucleosomes more diffuse) and, accordingly, in gene expression (77). In particular, it seems that upon methylation, the nucleosome-free region is less defined and the nucleosome −1 is moved downstream. Furthermore, our calculations suggest that when CpG islands are methylated, nucleosomes are concentrated at the CpG island edges, leading to a completely different configuration of the nucleosome array around the TSS and consequently to a change in gene activity. Hence, we can partially rationalize the striking effect of methylated CpG islands on the activity of several genes, particularly those involved in cancer (78), by considering the highly unfavorable impact that methylation has on the ability of poly CpG tracks to wrap around nucleosomes (Fig. 5 *A*). Further work is required to shed more light on this interesting hypothesis.

Taken together, our studies show how an apparently minor covalent change such as methylation can alter the physical properties of DNA, and how such a change can modify the ability of DNA to organize the chromatin fiber, which may be reflected by significant alterations in gene regulation, even in the absence of specific MBDs.

CONCLUSIONS

In summary, simple physical properties of DNA (described from calculations based on first principles) can provide a rationale for the seemingly chaotic diversity of gene regulatory signals in developed organisms, particularly epigenetic signatures such as cytosine methylation. Our results support the hypothesis that physical properties define a basal regulatory code that is superposed onto more elaborated

mechanisms involving the action of specific proteins when fine-tuning of gene function is required. Furthermore, our findings suggest that simply by varying the physical properties of some distant regions to a particular gene while keeping the specific protein-binding boxes unaltered, we may be able to modulate that gene's biological functionality. This raises interesting possibilities in the emerging field of synthetic biology. From the results of this study, it follows that methylated DNA is not as likely to form nucleosomes. However, the complete picture is even more complex when one considers that DNMTs have a greater preference to target nucleosome-bound DNA, slightly enriching it (1%) in ^{Me}CpG steps (79).

SUPPORTING MATERIAL

Supplementary materials and methods, seven figures, a table, and references are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(12)00397-9.

We thank Anna Aviñó for help with the experiments. The authors declare that they have no competing interests and acknowledge no conflict of interest.

This work was supported by the Spanish Ministry of Science and Innovation (BIO2009-10964 and Consolider E-Science, INB-Genoma España, and COMBIOMED RETICS), the European Research Council, and the Fundación Marcelino Botín. A.P. received support from a Juan de la Cierva postdoctoral fellowship.

REFERENCES

1. Collins, F. S., E. D. Green, …, M. S. Guyer, US National Human Genome Research Institute. 2003. A vision for the future of genomics research. *Nature.* 422:835–847.
2. Hélène, C. 1981. Recognition of base sequences by regulatory proteins in procaryotes and eucaryotes. *Biosci. Rep.* 1:477–483.
3. Schueler, M. G., and B. A. Sullivan. 2006. Structural and functional dynamics of human centromeric chromatin. *Annu. Rev. Genomics Hum. Genet.* 7:301–313.
4. Butcher, L. M., and S. Beck. 2008. Future impact of integrated high-throughput methylome analyses on human health and disease. *J. Genet. Genomics.* 35:391–401.
5. Bird, A. P. 1986. CpG-rich islands and the function of DNA methylation. *Nature.* 321:209–213.
6. McClelland, M., and R. Ivarie. 1982. Asymmetrical distribution of CpG in an 'average' mammalian gene. *Nucleic Acids Res.* 10:7865–7877.
7. Elango, N., and S. V. Yi. 2011. Functional relevance of CpG island length for regulation of gene expression. *Genetics.* 187:1077–1083.
8. Kundu, T. K., and M. R. Rao. 1999. CpG islands in chromatin organization and gene expression. *J. Biochem.* 125:217–222.
9. Doerfler, W. 1983. DNA methylation and gene activity. *Annu. Rev. Biochem.* 52:93–124.
10. Goll, M. G., and T. H. Bestor. 2005. Eukaryotic cytosine methyltransferases. *Annu. Rev. Biochem.* 74:481–514.
11. Leonhardt, H., A. W. Page, …, T. H. Bestor. 1992. A targeting sequence directs DNA methyltransferase to sites of DNA replication in mammalian nuclei. *Cell.* 71:865–873.
12. Clark, T. A., I. A. Murray, …, J. Korlach. 2012. Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. *Nucleic Acids Res.* 40:e29.
13. Klimasauskas, S., S. Kumar, …, X. Cheng. 1994. HhaI methyltransferase flips its target base out of the DNA helix. *Cell.* 76:357–369.
14. Chandler, L. A., and P. A. Jones. 1988. Hypomethylation of DNA in the regulation of gene expression. *Dev. Biol.* 5(N Y 1985):335–349.
15. Cedar, H., and Y. Bergman. 2009. Linking DNA methylation and histone modification: patterns and paradigms. *Nat. Rev. Genet.* 10: 295–304.
16. Southern, E. M. 1984. DNA sequences and chromosome structure. *J. Cell Sci. Suppl.* 1:31–41.
17. Suzuki, M. M., and A. Bird. 2008. DNA methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet.* 9:465–476.
18. Hermann, A., H. Gowher, and A. Jeltsch. 2004. Biochemistry and biology of mammalian DNA methyltransferases. *Cell. Mol. Life Sci.* 61:2571–2587.
19. Hermann, A., R. Goyal, and A. Jeltsch. 2004. The Dnmt1 DNA-(cytosine-C5)-methyltransferase methylates DNA processively with high preference for hemimethylated target sites. *J. Biol. Chem.* 279: 48350–48359.
20. Irizarry, R. A., C. Ladd-Acosta, …, A. P. Feinberg. 2009. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.* 41: 178–186.
21. Doi, A., I. H. Park, …, A. P. Feinberg. 2009. Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nat. Genet.* 41:1350–1353.
22. Reference deleted in proof.
23. Lister, R., M. Pelizzola, …, J. R. Ecker. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature.* 462:315–322.
24. Bhutani, N., J. J. Brady, …, H. M. Blau. 2010. Reprogramming towards pluripotency requires AID-dependent DNA demethylation. *Nature.* 463:1042–1047.
25. Reference deleted in proof.
26. Choi, S. H., K. Heo, …, A. S. Yang. 2011. Identification of preferential target sites for human DNA methyltransferases. *Nucleic Acids Res.* 39:104–118.
27. Lavery, R., K. Zakrzewska, …, J. Sponer. 2010. 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.* 38:299–313.
28. Pérez, A., I. Marchán, …, M. Orozco. 2007. Refinement of the AMBER force field for nucleic acids: improving the description of α/γ conformers. *Biophys. J.* 92:3817–3829.
29. Olson, W. K., A. A. Gorin, …, V. B. Zhurkin. 1998. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. USA.* 95:11163–11168.
30. Lankas, F., J. Sponer, …, T. E. Cheatham, 3rd. 2003. DNA basepair step deformability inferred from molecular dynamics simulations. *Biophys. J.* 85:2872–2883.
31. Morozov, A. V., K. Fortney, …, E. D. Siggia. 2009. Using DNA mechanics to predict in vitro nucleosome positions and formation energies. *Nucleic Acids Res.* 37:4707–4722.
32. Orozco, M., A. Noy, and A. Pérez. 2008. Recent advances in the study of nucleic acid flexibility by molecular dynamics. *Curr. Opin. Struct. Biol.* 18:185–193.
33. Pérez, A., F. Lankas, …, M. Orozco. 2008. Towards a molecular dynamics consensus view of B-DNA flexibility. *Nucleic Acids Res.* 36:2379–2394.
34. Lankas, F., J. Sponer, …, J. Langowski. 2000. Sequence-dependent elastic properties of DNA. *J. Mol. Biol.* 299:695–709.

35. Paillard, G., and R. Lavery. 2004. Analyzing protein-DNA recognition mechanisms. *Structure.* 12:113–122.
36. Pérez, A., F. J. Luque, and M. Orozco. 2012. Frontiers in molecular dynamics simulations of DNA. *Acc. Chem. Res.* 45:196–205.
37. Podtelezhnikov, A. A., C. Mao, …, A. Vologodskii. 2000. Multimerization-cyclization of DNA fragments as a method of conformational analysis. *Biophys. J.* 79:2692–2704.
38. Tolstorukov, M. Y., A. V. Colasanti, …, V. B. Zhurkin. 2007. A novel roll-and-slide mechanism of DNA folding in chromatin: implications for nucleosome positioning. *J. Mol. Biol.* 371:725–738.
39. Balasubramanian, S., F. Xu, and W. K. Olson. 2009. DNA sequence-directed organization of chromatin: structure-based computational analysis of nucleosome-binding sequences. *Biophys. J.* 96:2245–2260.
40. Thåström, A., P. T. Lowary, …, J. Widom. 1999. Sequence motifs and free energies of selected natural and non-natural nucleosome positioning DNA sequences. *J. Mol. Biol.* 288:213–229.
41. Deniz, O., O. Flores, …, M. Orozco. 2011. Physical properties of naked DNA influence nucleosome positioning and correlate with transcription start and termination sites in yeast. *BMC Genomics.* 12:489.
42. Anderson, J. D., and J. Widom. 2000. Sequence and position-dependence of the equilibrium accessibility of nucleosomal DNA target sites. *J. Mol. Biol.* 296:979–987.
43. Pérez, A., F. J. Luque, and M. Orozco. 2007. Dynamics of B-DNA on the microsecond time scale. *J. Am. Chem. Soc.* 129:14739–14745.
44. Behe, M., and G. Felsenfeld. 1981. Effects of methylation on a synthetic polynucleotide: the B—Z transition in poly(dG-m5dC).poly(dG-m5dC). *Proc. Natl. Acad. Sci. USA.* 78:1619–1623.
45. Banyay, M., and A. Gräslund. 2002. Structural effects of cytosine methylation on DNA sugar pucker studied by FTIR. *J. Mol. Biol.* 324: 667–676.
46. Schones, D. E., K. Cui, …, K. Zhao. 2008. Dynamic regulation of nucleosome positioning in the human genome. *Cell.* 132:887–898.
47. Kaplan, N., I. K. Moore, …, E. Segal. 2009. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature.* 458:362–366.
48. Chung, H. R., and M. Vingron. 2009. Sequence-dependent nucleosome positioning. *J. Mol. Biol.* 386:1411–1422.
49. Cui, F., and V. B. Zhurkin. 2010. Structure-based analysis of DNA sequence patterns guiding nucleosome positioning in vitro. *J. Biomol. Struct. Dyn.* 27:821–841.
50. Travers, A., E. Hiriart, …, E. Di Mauro. 2010. The DNA sequence-dependence of nucleosome positioning in vivo and in vitro. *J. Biomol. Struct. Dyn.* 27:713–724.
51. Trifonov, E. N. 2010. Nucleosome positioning by sequence, state of the art and apparent finale. *J. Biomol. Struct. Dyn.* 27:741–746.
52. Olson, W. K., and V. B. Zhurkin. 2011. Working the kinks out of nucleosomal DNA. *Curr. Opin. Struct. Biol.* 21:348–357.
53. Field, Y., N. Kaplan, …, E. Segal. 2008. Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLOS Comput. Biol.* 4:e1000216.
54. Bai, L., and A. V. Morozov. 2010. Gene regulation by nucleosome positioning. *Trends Genet.* 26:476–483.
55. Segal, E., Y. Fondufe-Mittendorf, …, J. Widom. 2006. A genomic code for nucleosome positioning. *Nature.* 442:772–778.
56. Lee, W., D. Tillo, …, C. Nislow. 2007. A high-resolution atlas of nucleosome occupancy in yeast. *Nat. Genet.* 39:1235–1244.
57. Tillo, D., and T. R. Hughes. 2009. G+C content dominates intrinsic nucleosome occupancy. *BMC Bioinformatics.* 10:442.
58. Goñi, J. R., C. Fenollosa, …, M. Orozco. 2008. DNAlive: a tool for the physical analysis of DNA at the genomic scale. *Bioinformatics.* 24:1731–1732.
59. Goñi, J. R., A. Pérez, …, M. Orozco. 2007. Determining promoter location based on DNA structure first-principles calculations. *Genome Biol.* 8:R263.
60. Bertrand, H., T. Ha-Duong, …, B. Hartmann. 1998. Flexibility of the B-DNA backbone: effects of local and neighbouring sequences on pyrimidine-purine steps. *Nucleic Acids Res.* 26:1261–1267.
61. Nathan, D., and D. M. Crothers. 2002. Bending and flexibility of methylated and unmethylated EcoRI DNA. *J. Mol. Biol.* 316:7–17.
62. Mirsaidov, U., W. Timp, …, G. Timp. 2009. Nanoelectromechanics of methylated DNA in a synthetic nanopore. *Biophys. J.* 96:L32–L34.
63. Bell, A. C., and G. Felsenfeld. 2000. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature.* 405:482–485.
64. Clark, S. J., J. Harrison, and P. L. Molloy. 1997. Sp1 binding is inhibited by (m)Cp(m)CpG methylation. *Gene.* 195:67–71.
65. Hark, A. T., C. J. Schoenherr, …, S. M. Tilghman. 2000. CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature.* 405:486–489.
66. Davey, C. S., S. Pennings, …, J. Allan. 2004. A determining influence for CpG dinucleotides on nucleosome positioning in vitro. *Nucleic Acids Res.* 32:4322–4331.
67. Pennings, S., J. Allan, and C. S. Davey. 2005. DNA methylation, nucleosome formation and positioning. *Brief. Funct. Genomics Proteomics.* 3:351–361.
68. Felle, M., H. Hoffmeister, …, G. Längst. 2011. Nucleosomes protect DNA from DNA methylation in vivo and in vitro. *Nucleic Acids Res.* 39:6956–6969.
69. Choy, J. S., S. Wei, …, T. H. Lee. 2010. DNA methylation increases nucleosome compaction and rigidity. *J. Am. Chem. Soc.* 132:1782–1783.
70. Lee, J. Y., and T. H. Lee. 2012. Effects of DNA methylation on the structure of nucleosomes. *J. Am. Chem. Soc.* 134:173–175.
71. Lee, C. K., Y. Shibata, …, J. D. Lieb. 2004. Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nat. Genet.* 36: 900–905.
72. Field, Y., Y. Fondufe-Mittendorf, …, E. Segal. 2009. Gene expression divergence in yeast is coupled to evolution of DNA-encoded nucleosome organization. *Nat. Genet.* 41:438–445.
73. Choi, J. K., J. B. Bae, …, Y. J. Kim. 2009. Nucleosome deposition and DNA methylation at coding region boundaries. *Genome Biol.* 10:R89.
74. Feng, J., X. Dai, …, C. He. 2010. New insights into two distinct nucleosome distributions: comparison of cross-platform positioning datasets in the yeast genome. *BMC Genomics.* 11:33.
75. Kaplan, N., I. Moore, …, E. Segal. 2010. Nucleosome sequence preferences influence in vivo nucleosome organization. *Nat. Struct. Mol. Biol.* 17:918–920, author reply 920–912.
76. Valouev, A., J. Ichikawa, …, S. M. Johnson. 2008. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Res.* 18:1051–1063.
77. Wang, X., G. O. Bryant, …, M. Ptashne. 2011. An effect of DNA sequence on nucleosome occupancy and removal. *Nat. Struct. Mol. Biol.* 18:507–509.
78. Esteller, M. 2006. Epigenetics provides a new generation of oncogenes and tumour-suppressor genes. *Br. J. Cancer.* 94:179–183.
79. Chodavarapu, R. K., S. Feng, …, M. Pellegrini. 2010. Relationship between nucleosome positioning and DNA methylation. *Nature.* 466:388–392.

1.3. Unraveling the hidden DNA structural/physical code provides novel insights on promoter location

ProStar is a promoter predictor from DNA primary sequence using only physical descriptors (Goñi *et al.*, 2007). Despite being developed few years ago, in this work its application is proven by discovering regulatory activity in regions without annotated genes previously annotated as false positives.

In this work, a set of putative core promoters sequences without known TSSs annotated predicted by ProStar were subjected to different transcriptional analysis, including luciferase assays, CAGE and RNA-seq. Obtained results showed a significant transcriptional activity in the positive set of predictions in comparison with control set. The fact that many of those predicted sequences lack known specific sequence motifs concludes that the predictive power of physical signals goes beyond the primary sequence.

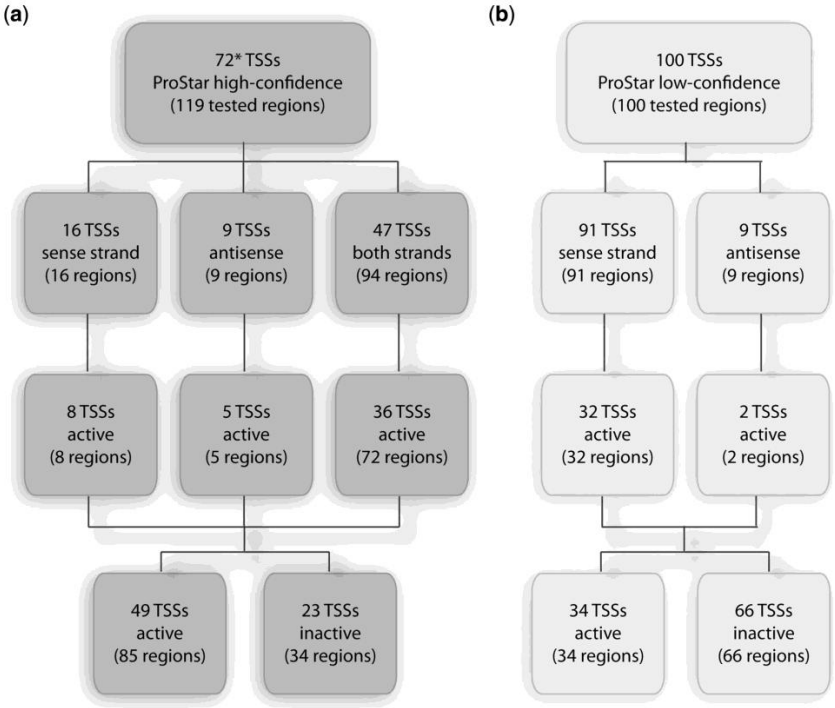


Fig. 20 Summary scheme of TSS selection for both positive (a) and negative (b) ProStar sets, classified based on luciferase activity (3-fold) and directionality of tested regions: sense, antisense or both sense strands. (Asterisk) 24 out of 72 TSSs were annotated on recent transcriptome reference annotations based on the 2009 genome release (GRChR37/hg19), i.e. they are true positives.

Publication:

Duran, E., Djebali, S., Gonzalez, S., Flores, O., Mercader, J. M., Guigó, R., Torrents, D. Soler-López, M., Orozco, M. (2013) "*Unravelling the hidden DNA structural/physical code provides novel insights on promoter location*", Nucleic Acids Research (early access) doi: 10.1093/nar/gkt511

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

Nucleic Acids Research Advance Access published June 12, 2013

Nucleic Acids Research, 2013, 1–11
doi:10.1093/nar/gkt511

Unravelling the hidden DNA structural/physical code provides novel insights on promoter location

Elisa Durán[1,2], Sarah Djebali[3], Santi González[2,4], Oscar Flores[1,2], Josep Maria Mercader[2,4], Roderic Guigó[3], David Torrents[2,4], Montserrat Soler-López[1,2] and Modesto Orozco[1,2,4,5,*]

[1]Institute for Research in Biomedicine (IRB Barcelona), Barcelona 08028, Spain, [2]Joint IRB-BSC Research Program on Computational Biology, Barcelona 08028, Spain, [3]Bioinformatics and Genomics Group, Center for Genomic Regulation and Universitat Pompeu Fabra, Barcelona 08003, Spain, [4]Barcelona Supercomputing Center, Barcelona 08034, Spain and [5]Department of Biochemistry and Molecular Biology, University of Barcelona, Barcelona 08028, Spain

Received October 23, 2012; Revised February 15, 2013; Accepted April 30, 2013

ABSTRACT

Although protein recognition of DNA motifs in promoter regions has been traditionally considered as a critical regulatory element in transcription, the location of promoters, and in particular transcription start sites (TSSs), still remains a challenge. Here we perform a comprehensive analysis of putative core promoter sequences relative to non-annotated predicted TSSs along the human genome, which were defined by distinct DNA physical properties implemented in our ProStar computational algorithm. A representative sampling of predicted regions was subjected to extensive experimental validation and analyses. Interestingly, the vast majority proved to be transcriptionally active despite the lack of specific sequence motifs, indicating that physical signaling is indeed able to detect promoter activity beyond conventional TSS prediction methods. Furthermore, highly active regions displayed typical chromatin features associated to promoters of housekeeping genes. Our results enable to redefine the promoter signatures and analyze the diversity, evolutionary conservation and dynamic regulation of human core promoters at large-scale. Moreover, the present study strongly supports the hypothesis of an ancient regulatory mechanism encoded by the intrinsic physical properties of the DNA that may contribute to the complexity of transcription regulation in the human genome.

INTRODUCTION

Gene expression in eukaryotes is a complex process regulated by a myriad of molecular mechanisms. The protein recognition of specific DNA sequence motifs located on promoter regions, upstream of transcription start sites (TSSs), has been traditionally considered as the most important regulatory element in transcription (1,2). Nevertheless, after one decade of the postgenomic era, the location of promoters and in particular TSSs still remains surprisingly challenging (3–6). Classical assumptions such as their location 5′ upstream of transcribed regions or their one-to-one correlation with coding genes might actually be oversimplistic. Indeed, sequence signals like transcription factor–binding sites (TFBSs) show little predictive power when applied at the entire genome level. Furthermore, massive annotation projects (7–9) have provided further evidence about the complexity of promoter location and its occurrence in rather unusual genomic regions. These difficulties illustrate that the mechanisms regulating gene expression are not exclusively based on specific interactions between nucleobases located upstream TSSs and regulatory proteins, as they would lead to detectable sequence signals otherwise. Conversely, it seems that the world of DNA regulation is much more intricate and probably involves a myriad of mechanisms, such as the modulation of chromatin structure or epigenetic signatures (10,11).

We and others (12–15) have suggested the existence of a physical code imprinted onto the DNA fiber, which could account for an ancient regulatory mechanism of basal gene expression. Indeed, core promoters and associated TSSs are DNA segments with an intrinsic ability to act as regulatory regions, as they are depleted in nucleosomes and need to bind to a large number of regulatory proteins, which certainly require special physical properties of the DNA fiber. According to this paradigm, we consider that promoters can be defined as regions of unusual physical deformability (13,15,16), which (even in the absence of traditional sequence motifs) might favor either a suitable nucleosome positioning pattern for protein recognition

*To whom correspondence should be addressed. Tel: +34 934037155; Fax: +34 934037157; Email: modesto.orozco@irbbarcelona.org

(17) or an effective binding of core promoter-binding proteins and RNA polymerase (12,18). Notwithstanding, genome-wide analysis of the DNA physical properties (13) revealed that 'promoter-like' physical signals appear in regions without evidence for real promoters, challenging the existence of a regulatory physical code in DNA, or alternatively, suggesting the presence of many hidden promoter regions in the human genome.

In this manuscript, we have revisited our presumptions about the existence of a physical code involved in gene activity regulation. To this end, we have evaluated de novo promoter predictions arising from the location of regions with unusual physical properties (13). A representative set of suggested (but not annotated) promoters have been analyzed by applying a combination of medium and high-throughput experimental techniques and analyses. Our study demonstrates that a strikingly large number of theoretical predictions, which were considered 'false positives' based on the 2007 knowledge, are indeed true promoters. Therefore, we have been able to determine many novel TSSs and core promoters, which were neither detectable by alternative methods nor presenting orthologous sequence signals with known promoters. Most importantly, the present study enables us to redefine promoter signatures and analyze the diversity, evolutionary conservation and dynamic regulation of human core promoters at large-scale. Overall, our findings provide a solid support to the hypothesis that a primitive physical code imprinted in the DNA fiber constitutes a first level of regulation of gene activity.

MATERIALS AND METHODS

ProStar promoter predictor

Our ProStar promoter prediction program is able to predict TSSs based on the presence of an unusual profile of physical properties (particularly the DNA helical stiffness) (13), simplifying previous algorithms that use a variety of empirical descriptors with complex translation to mechanistic models (12). As described elsewhere (13,15), stiffness parameters were derived from atomistic molecular dynamics simulations using model oligonucleotides, annotated at the dinucleotide level, and averaged linearly along 500 bp size windows. In short, we performed a large number of molecular dynamics (MD) simulations, computing then the covariance matrices in the helical space at the dinucleotide step $[d(X \cdot Y)/d(Z \cdot T)]$. Inversion of such matrix yields a 6×6 stiffness matrix for each dinucleotide step (13,15). To keep the model as simple as possible, we only considered diagonal elements of the matrix, i.e. the stiffness of DNA in front of pure 'twist', 'roll', 'tilt', 'rise', 'slide' and 'shift' deformations. The average physical property profiles were defined from the analysis of two genomic sequence sets (NCBI36/hg18 human genome release, March 2006), corresponding to known promoters (positive set) or randomly selected sequences (negative set) according to the reference GENCODE annotation (19). ProStar scores a given DNA sequence as 'promoter' or as 'background' depending on its similarity to the two reference profiles. This is computationally measured by the Mahalanobis distance—a simple statistical metrics widely implemented in clustering and classification analyses (20)—to both promoter and background reference profiles. Using ProStar default parameters, 500 bp long DNA sequences were analyzed at the genome-wide level to locate potential TSSs (13). In this work, putative human core promoters were identified as regions within a window of −1000/+200 bp relative to the ProStar-predicted TSS locations.

Selection of TSS prediction sets

To be coherent with the ProStar training, we applied our predictor using ENSEMBL (v47) (21) as a reference annotation to select TSSs located at least 1200 bp away from any other annotated TSS. As a result, we obtained a set of putative 'false positive', i.e. regions predicted as promoters by their unusual physical properties but which were not experimentally known. We then filtered out those regions that presented >70% of repetitive elements according to the RepeatMasker algorithm (http://www.repeatmasker.org), or that did not allow unique polymerase chain reaction (PCR) primer localization to the human genome assembly by *in silico* PCR BLAT search (http://genome.ucsc.edu). This process yielded 119 genomic regions (1200 bp long) located around 72 putative TSS (note that it was not always technically possible to study promoters located in both directions).

As a negative prediction set, we randomly selected 100 positions, where ProStar suggested no TSS in a 1200 bp window, and for which unique PCR primers could be located. To make the test unbiased, we did not perform any filtering based on the presence of 2006 known promoters in these ProStar negative predictions. Both ProStar-positive and ProStar-negative predicted promoters were subjected to experimental validation.

The positive set was further compared against the latest gene and transcript reference annotations GENCODE (v7) (19) and ENSEMBL (v56) (21) to determine the true positives.

Luciferase transcription activity assays

We designed hybridization primers suitable for high-GC content regions. The presence of a unique hybridization site was subsequently verified by a BLAT genome alignment (http://genome.ucsc.edu). Primers were ordered in 96-well plates to Sigma-Aldrich. PCR was performed in a 96-well format using AccuPrime GC-rich DNA polymerase (Invitrogen) for the amplification of selected regions. PCR products were analyzed in a 1% agarose gel. Successfully amplified regions were inserted into the promoterless pGL4.21 (luc2P/Puro) vector and ligated through Sfi I restriction sites (Rapid DNA ligation Kit, Roche) that enable directional cloning. *Escherichia coli* competent cells (DH5α, Invitrogen) were transformed with the ligation products. Two independent colonies were selected from each transformant and were verified by sequencing from both the 5′ and 3′ ends. The experimental approach for luciferase activity assays in a high-throughput approach is outlined in Supplementary Figure S1.

Cos-7, Hek293, U2OS, MIA PACA and MDA231 cells were cultured in Dulbeccós Modified Eaglés Media (DMEM) supplemented with 10% of fetal calf serum (FBS). All cultures were grown as a monolayer in a humidified incubator at 37°C in an atmosphere of 5% CO_2. One day before co-transfection, $2–6 \times 10^4$ cells per well were plated in 96-well plates with 100 µl of DMEM without antibiotics. Confluence of 90–95% was achieved by the second day. Transient DNA co-transfections were performed with 0.1 µg of the corresponding pGL4.21/construct plasmid and 0.02 µg of the pGL4.74 (*hRluc*/*TK*) vector (Promega) using TransFact reagent (Promega) according to the instructions of the manufacturer. DMEM supplemented with 10% FBS was added to the cells 1 h after co-transfection to allow correct growth and protein expression. Dual Luciferase Reporter Assay (Promega) was performed 36 h after co-transfection using a GloMax Multidetection Luminometer (Promega) with dual injector system allowing rapid reagent addition. Light emission was measured 2 s after addition of each of the substrates and integrated over a 10-s interval. The firefly luciferase activity results were normalized with the renilla luciferase activity from the pGL4.74 (hRluc/TK) plasmid to account for differences in transfection efficiency. The previously characterized *SPG4* gene promoter (22) was used to generate positive (S−621/−1) and negative (S−1290/−424) promoter region controls, respectively. Promoter activity was assessed in duplicates and was considered active if it exceeded 3-fold the score of negative control sequences from the normalized threshold value.

After luciferase assays, 80 regions from both the positive and negative promoter sets were further divided into four subsets for further analysis: subset 1 contains 20 high-confidence ProStar sequences with high luciferase activity (PS+L+); subset 2 contains another 20 high-confidence ProStar sequences with low luciferase activity (PS+L−); subset 3, 20 low-scored ProStar sequences with luciferase activity (PS−,L+); and subset 4, 20 low-scored ProStar sequences with no luciferase activity (PS−L−).

CAGE analysis

To measure transcription initiation in the different promoter subset regions, profiles of cap analysis gene expression (CAGE) 5′-ends were computed. For this purpose, ENCODE stranded CAGE data from polyadenylated cytosolic RNA of seven different cell lines (GM12878, H1-hESC, HUVEC, HeLa-S3, HepG2, K562 and NHEK) and generated in two bio-replicates were used (23–25). For each cell line, CAGE mappings of quality >20 from each of the two bio-replicates were merged, and their distinct 5′-ends extracted (redundancy was removed to avoid considering reverse transcriptase-PCR artifacts as true signal). Every region was subjected to two CAGE analyses, either considering the luciferase-tested 1200 bp region or a 2000 bp equivalent expanded region centered at the TSS. For every cell line, each time a CAGE tag 5′-end was located within and on the same strand as one of the promoter regions, the distance between the CAGE tag 5′-end and the promoter region 5′-end was computed, and the CAGE frequency corresponding to this distance (further normalized using percentage distance bins) was increased.

RNA-seq analysis

To measure transcription activity in the different promoter subset regions, profiles of RNA-seq 5′-ends were computed. For this purpose, ENCODE CSHL stranded paired-end RNA-seq data from polyadenylated cytosolic RNA of seven different cell lines (GM12878, H1-hESC, HUVEC, HeLa-S3, HepG2, K562 and NHEK) were used (25). For each cell line, all the mappings of the second bio-replicate were considered, and their distinct most 5′-ends extracted. Every subset region was expanded to a final length of 2000 bp centered at the TSS, similarly to the CAGE analyzed sequences. For every cell line, each time an RNA-seq mapping 5′-end was located within and on the same strand as one of the promoter regions, the distance between the RNA-seq 5′-end and the promoter region 5′-end was computed, and the RNA-seq frequency corresponding to this distance (further normalized using percent distance bins) was increased.

Chromatin structure and epigenetic signals

The chromatin structure was inferred from DNase I hypersensitivity sites as reported in ENCODE through the UCSC Table Browser data retrieval tool (26). From these data, we calculated the average of Dnase I hypersensitiviy clusters within 1200 bp regions of the different CAGE analyzed subsets, considering a positive cluster when overlapped with the reference promoter elements. We also explored potential epigenetic markers in the suggested promoter regions by looking at the occurrence of histone variants H3KMe1, H3K27Ac and H3K4Me3 in seven different cell lines (GM, H1, HSMM, HUVEC, K562, NHEK and NHLF). For each CAGE analyzed subset of 1200 bp regions, we calculated the number of regions that overcome a certain average alignment density (intensity signal) in any of the different cell types. Using a threshold of 10-fold, 92% of PS+ sequences contained stronger signals compared with the 37% of PS−. Increasing the threshold, up to 50, produced a reduction of the total number of regions, but increased the difference between PS+ and PS− in the same direction.

TFBS enrichment evaluation

We investigated if different subsets, including the PS+ predictions (17 909 in total), the experimentally tested PS+ predictions (119 sequences) and PS− predictions (100 sequences) or luciferase positive (49 sequences) and negative (23 sequences) regions, were enriched within the 1200 bp in any of the currently annotated 885 TFBSs. To this end, we systematically compared them with a full list of transcripts described in the BioMart database (http://www.biomart.org) (76 905 transcripts) as a background control. To determine the significant enrichment, we used a Fisher's exact test and represented the magnitude of enrichment as odds ratios, which is the ratio of enrichment for a given TFBS. The corrected significant *P*-value after applying a Bonferroni's correction for all tests was $0.05/885 = 5.65 \times 10^{-5}$. The analyses were performed using the R statistical environment (http://www.r-project.org).

Core region DNA element conservation and sequence-based signals

The conservation of the four different CAGE analyzed 1200 bp regions was evaluated by the comparison with available vertebrate genomes using the University of California, Santa Cruz (UCSC) Table Browser data retrieval tool (26). The level of conservation for each particular fragment was calculated according to Vertebrate Basewise Conservation by PhyloP as the average of conservation of all nucleotides comprising the region. TFBS conservation was determined from the comparison of boxes among human, mouse and rat according to the UCSC TFBS conservation track using matrices obtained from TRANSFAC database (27). In addition, we used Regulatory region Local Alignment (ReLA) algorithm (28), a footprinting-based program for the detection of conserved clusters of TFBSs, to determine whether the regions predicted by ProStar would also be detectable as sequence-based only promoter signals.

RESULTS

Selection of the TSS prediction sets based on DNA physical properties

If physical signals were indeed significant in regulatory regions, as we presume, we would expect a high proportion of ProStar predictions to be promoters, despite the lack of experimental annotation. As described elsewhere (13), promoter sequences provide a distinct profile for six descriptors of the DNA stiffness in front of 'twist', 'roll', 'tilt', 'rise', 'slide' and 'shift' deformations, particularly in regions spanning −250/+900 bp relative to TSSs (that is, covering core and proximal promoter distances).

Therefore, to validate our hypothesis, we first defined a TSS prediction set for experimental screening from the ProStar genome-wide calculations. To better validate the prediction power, we used the original ProStar outcome based on the 2006 release of the human genome (13), without retraining the software with more recent releases of genome data. We selected regions with unusual physical properties suggested to be promoters albeit they were not annotated in reference databases, i.e. ProStar 'false positives' (PS+, see 'Materials and Methods' section). Furthermore, even though the algorithm recognizes the directionality of transcription, predictions might also account for bidirectional regulatory elements. Thereby, we selected 72 high-scored putative TSSs that allowed unique PCR primer hybridization to the human genome assembly on the sense-strand (16 TSSs), antisense (9 TSSs) or in both senses (47×2 TSSs), yielding 119 different putative promoter regions in total (Figure 1a, Supplementary Table S1). We additionally defined a negative set consisting of 100 sequences corresponding to nonsignaled promoter regions by ProStar (PS−, 91 on direct-sense and 9 anti-sense due to PCR constraints) (Figure 1b, Supplementary Table S1).

Comparison of the physical deformability properties between both sets revealed the distinct underlying features that had allowed ProStar to recognize the positive TSS set as putative promoter regions, as described above (15). When we further compared our positive set against the latest transcript reference annotations GENCODE (v7) (19) and ENSEMBL (v56) (21), 24 predicted TSSs appeared to be functional (i.e. they are certainly true positives), giving an unexpected support to the quality of our physical de novo predictions (Supplementary Table S1). Yet, up to 48 ProStar predicted regions are not proximate (<1.2 kb) to any 2012-annotated TSS. Intriguingly, the attempt of validating ProStar predicted regions using methods based on interspecies sequence conservation, such as ReLA (28), yielded a low success rate (9%), providing further evidences that ProStar locates putative promoters in genomic regions where phylogenetic footprinting finds no signal.

Identification of functional promoters

We evaluated the ability of the selected putative regions to activate transcription in mammalian cells by using luciferase reporter gene expression assays (Supplementary Figure S1). By applying a threshold of at least 3-fold higher activity than the control vehicle, 85 putative promoters regions were scored as functional, with a validation rate of 71.4%, while only 34% of the analyzed regions in the negative set displayed activity (Figure 1; Supplementary Table S1). From those 85 positively active sequences, 8 correspond to sense strand, 5 to antisense and 72 to both directions (i.e. putative bidirectional promoters or alternative regulatory elements), accounting for 49 distinct TSSs. Interestingly, a significantly large number of the suggested promoters (37.8%) displayed high activity (10-fold above the vehicle). Furthermore, almost 98% of active promoters in one cell line also displayed activity in three additional cell lines, indicating that the identified regions would mainly generate transcripts involved in housekeeping activities, rather than in tissue-specific processes. Taken together, these findings suggest that physical properties would signal promoters of the loosely regulated 'housekeeping' genes, whereas highly specific sequence signals would be required for the activation of development or tissue-specific genes.

CAGE and RNA-Seq analyses in support of predicted TSSs

Luciferase measurements showed that the vast majority of ProStar TSS-derived regions function as promoters when coupled to a reporter gene and transfected to mammalian cells (Figure 1; Supplementary Table S1). Nevertheless, we should also consider that the resulting activity could be an artifact for some regions, as the activity measurements were based on plasmid-inserted regions rather than on their native structure like in bulk chromatin. Alternatively, the activity might result from the absence of methylation or other posttranslational modifications of a true cellular environment, which can modify the DNA physical properties and ultimately lead to a transcription repression *in vivo* (29–31).

Consequently, we complemented our first validation with a CAGE (7,32,33) to examine the transcription start activity of the experimentally tested 1200 bp regions in living cells (25). We selected 80 regions showing

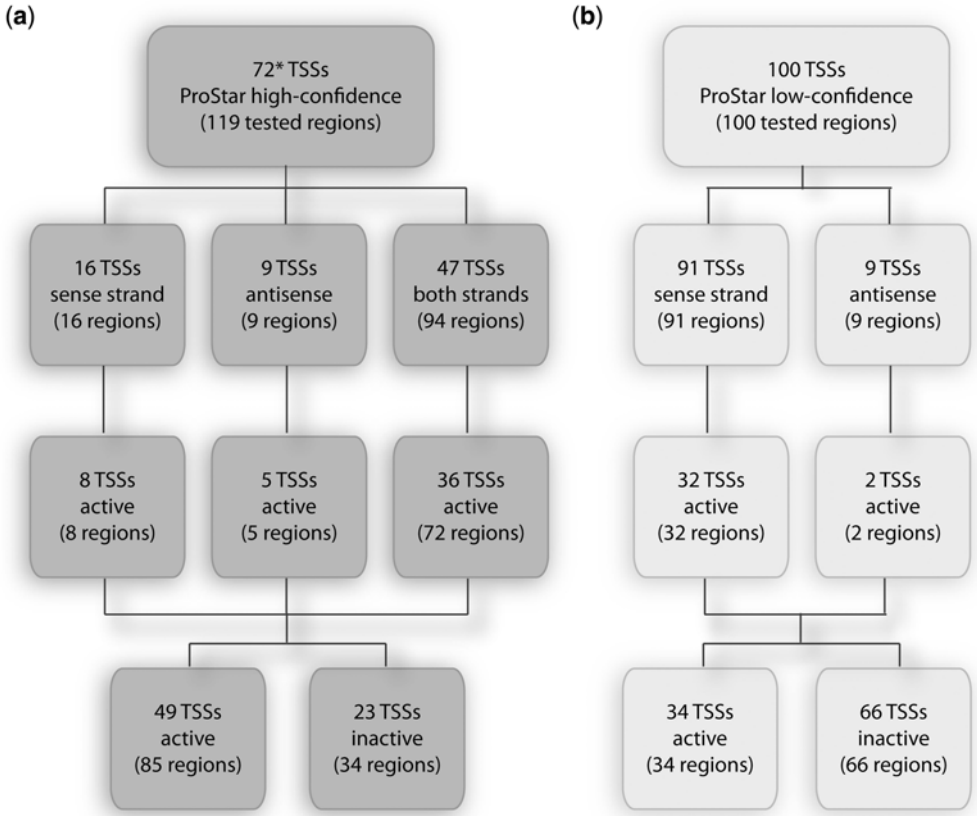


Figure 1. Identification of functional promoters. Summary scheme of TSS selection for both positive (**a**) and negative (**b**) ProStar sets, classified based on luciferase activity (3-fold) and directionality of tested regions: sense, antisense or both sense strands. (Asterisk) 24 out of 72 TSSs were annotated on recent transcriptome reference annotations (21) based on the 2009 genome release (GRChR37/hg19), i.e. true positives.

different levels of luciferase activity and classified them into four representative categories. Subset 1 contains 20 ProStar high-scored sequences with high luciferase activity (PS+L+); subset 2 contains another 20 ProStar high-scored sequences with low luciferase activity (PS+L−); subset 3, 20 low-scored ProStar sequences with luciferase activity (PS−L+); and subset 4, 20 low-scored ProStar sequences with no luciferase activity (PS−L−) (Supplementary Table S1).

The results summarized in Figure 2 show that regions from subsets 1 (PS+L+) and 2 (PS+L−) were dramatically enriched for CAGE tags that could be confidently mapped to single positions (Figure 2a and b), as compared with the ProStar negative subsets 3 (PS−L+) and 4 (PS−L−) (Figure 2c and d). Subset 1 displayed the highest proportion of sequences with CAGE tagged 5′-ends around 750 bp, indicating that those regions contained reliable TSS marks (Figure 2a, around 60th distance bin). CAGE tags were detected in most of the human cell type experiments, but a particular enrichment was found for polyadenylated (polyA+) transcripts, suggesting that active regions might correspond to promoter elements regulating protein-coding genes.

Interestingly, subset 3 (PS−L+) regions contain few cage tags, although they showed some activity in luciferase expression assays (Figure 2c). We could simply assume that this subset contains luciferase-false positives. However, it has been reported that the structure of promoters on different chromosomes varies and these variations might not be well covered by whole-genome promoter prediction algorithms (6). Thus, we cannot rule out the possibility that promoters located in anomalous positions, and hence harboring a divergent pattern of physical properties, could have been overlooked by ProStar (13,15). If these regulatory elements turn out to be under tight regulation in bulk chromatin (which would explain why no CAGE tags are detected), they could well show transcriptional activity in luciferase assays, which ignore activation or inhibition signals imprinted in the native chromatin structure.

Even more intriguingly, subset 2 regions (PS+L−) did show clear CAGE enrichment although they did not provide a luciferase response (Figure 2b). These discrepancies could simply result from luciferase-false negatives. However, the strength and the profile of CAGE signals (Figure 2b) indicated that other factors could also account for the low luciferase/high CAGE response. Comparison of the CAGE profiles indicated that subset 1 peaks are located at the expected TSSs (i.e. around 60th bin; Figure 2a), while subset 2 peaks are upstreamly

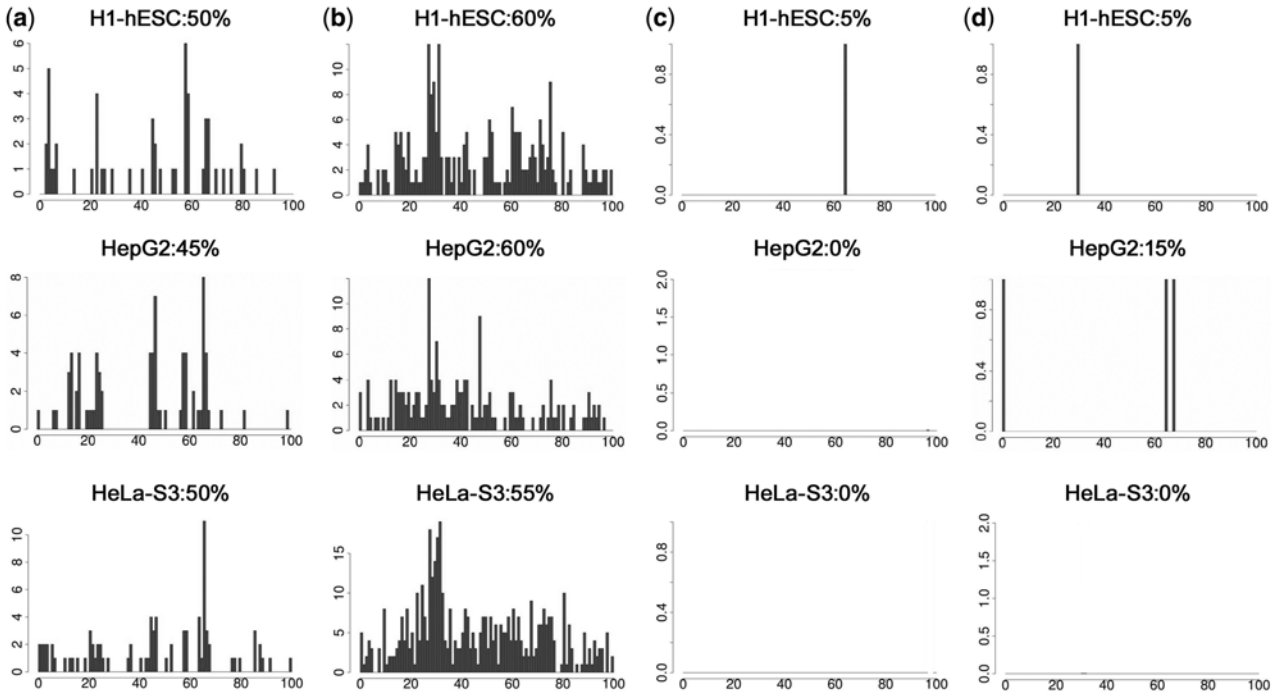


Figure 2. Orthogonal support of predicted TSSs: CAGE analysis. Distribution of distinct 5′-ends of CAGE tags from several representative CAGE experiments in H1-hESC, HepG2 and HeLa-S3 cell types based on cytosolic polyA+ transcripts. For every distinct most 5′-end of CAGE tag detected within and on the same strand as a particular promoter region, we increased the CAGE frequency of the percent distance bin corresponding to the distance between the CAGE tag 5′-end and the promoter region 5′-end. As the predicted promoter regions were 1200 bp long, each % distance bin includes 12 bp, and thereby the TSS is expected to be located on the 84th distance bin (i.e. at 1000 bp from the region 5′-end). (**a**) PS+L+ subset 1. For most of the cell types, the major peak appears around the 63th bin (i.e. 750 bp), closely matching with the prediction (**b**) PS+L− subset 2. We observe undefined peaks around the 30th–50th bins (350–600 bp). On the other hand, the number of CAGE tags is significantly higher than for subset 1 (**c**) PS−L+ for subset 3, (**d**) PS−L− for subset 4. ProStar negative PS− subsets clearly show an almost inexistent CAGE signal.

displaced from the original prediction (around 30th bin; Figure 2b). These findings suggest that, under certain conditions, physical properties are able to signal promoter regions although the prediction of the TSS location can be upstreamly displaced from the true site. In this scenario, CAGE experiments would still detect transcript 5′ in the −1000/+200 bp analyzed genomic window. On the other hand, this displacement would have led us to amplify truncated promoter constructs undetectable by the conservative luciferase test we initially applied in our experimental workflow (Supplementary Figure S1).

To validate this hypothesis, we carried out RNA-sequencing (RNA-seq) analysis to survey the transcription profiles of the selected regions and to identify putative exons near the suggested TSSs. We performed the analysis of 2000 bp regions centered on the predicted TSSs, using RNA-seq data of subcellular-fractionated RNAs from the ENCODE Consortium (Supplementary Figure S2, see 'Materials and Methods' section for details) (9,25). Interestingly, the profiles of subset 1 presented a sharp RNA-seq peak at 800 bp, which coincided with the CAGE major peak around 750 bp (Figure 3a, around 40th bin, orange frames). Furthermore, this TSS putative peak was corroborated with a downstream peak corresponding to an exon (50–80th bins, i.e. from 1000 to 1400 bp) in most of the cell lines. Conversely, subset 2 profiles showed two sharp RNA-seq peaks at 200 and 400 bp, respectively, which matched CAGE major peaks around 360 bp (Figure 3b, 10–20th bins, highlighted with orange frames). Moreover, a downstream broad peak likely corresponding to an exon (Figure 3b, 20–40th bins, i.e. from 400 to 800 bp, highlighted with purple frames) could further confirm the TSS displaced positions at ∼700–500 bp upstream relative to predictions.

We further interrogated this potential TSS displacement in the prediction by analyzing new genomic fragments but now centered on the observed CAGE peaks. To this end, we picked up regions from subset 2 and placed the TSS 500 bp upstream to the original ProStar TSS prediction, as indicated by the CAGE/RNA-seq profiles (Figure 4a). As expected, CAGE profiles exhibited a major peak around 800–900 bp, resembling subset 1 sequences (Figure 4b, around 45th bin). Similarly, RNA-seq profiles also presented a single peak at the expected position (Figure 4c, around 50th bin, 1000 bp). We then re-amplified four of these genomic regions by PCR, spanning 2000 bp but centered at the newly located TSS, as similarly done with previous subsets (Figure 4a; see Supplementary Figure S1 for method details). Interestingly, luciferase assays measured a 4-fold higher activity on average than the original sequences (Figure 4d), providing further evidence that subset 2 segments (PS+L−) do contain true TSSs.

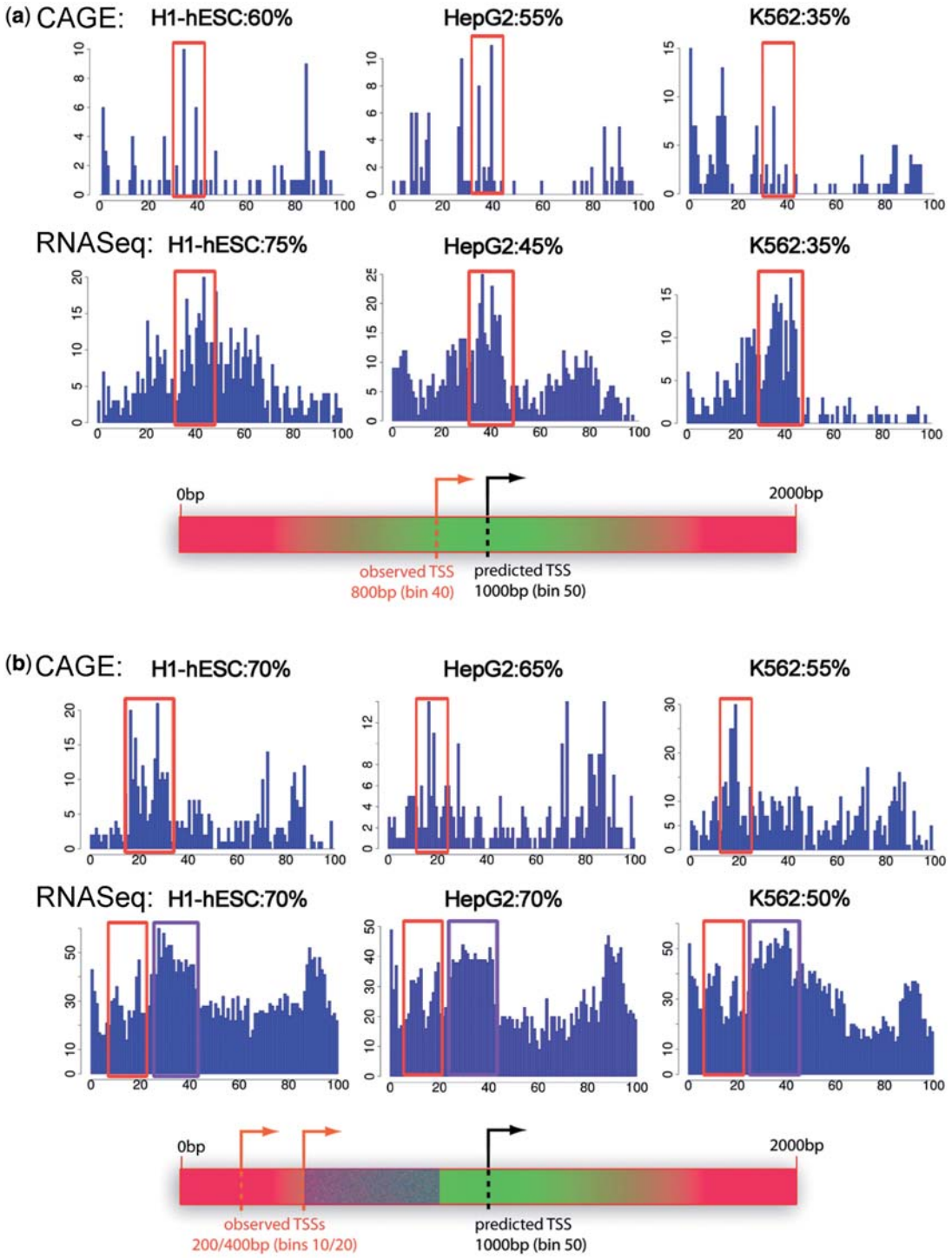


Figure 3. Orthogonal support of predicted TSSs: CAGE vs RNA-seq analyses. Distribution of 5′-ends of CAGE/RNA-seq tags from representative CAGE/RNA-seq experiments in H1-hESC, HepG2 and K562 cell types based on cytosolic polyA+ transcripts. The profiles were constructed similarly to the 1200 bp CAGE analysis. However, as the predicted promoter regions were now 2000 bp long, each % distance bin includes 20 bp and hence the predicted TSS should be located on the 50th distance bin (i.e. 1000 bp), as it is indicated in the promoter region schematic representations below the profiles (**a**) PS+L+ subset 1. The observed TSSs extrapolated from CAGE and RNA-seq profiles appear around the 40th bin (i.e. 800 bp, highlighted with orange frames), and closely match the predictions (**b**) PS+L− subset 2. We observe two sharp RNA-seq peaks around the 10th–20th bins (200–400 bp) that match with CAGE peaks around the 20th bin (highlighted with orange frames). Furthermore, a broad peak is observed right after the observed TSSs, indicating that it may correspond to a transcription active region (i.e. an exon, highlighted in purple) but not necessarily a transcription start region.

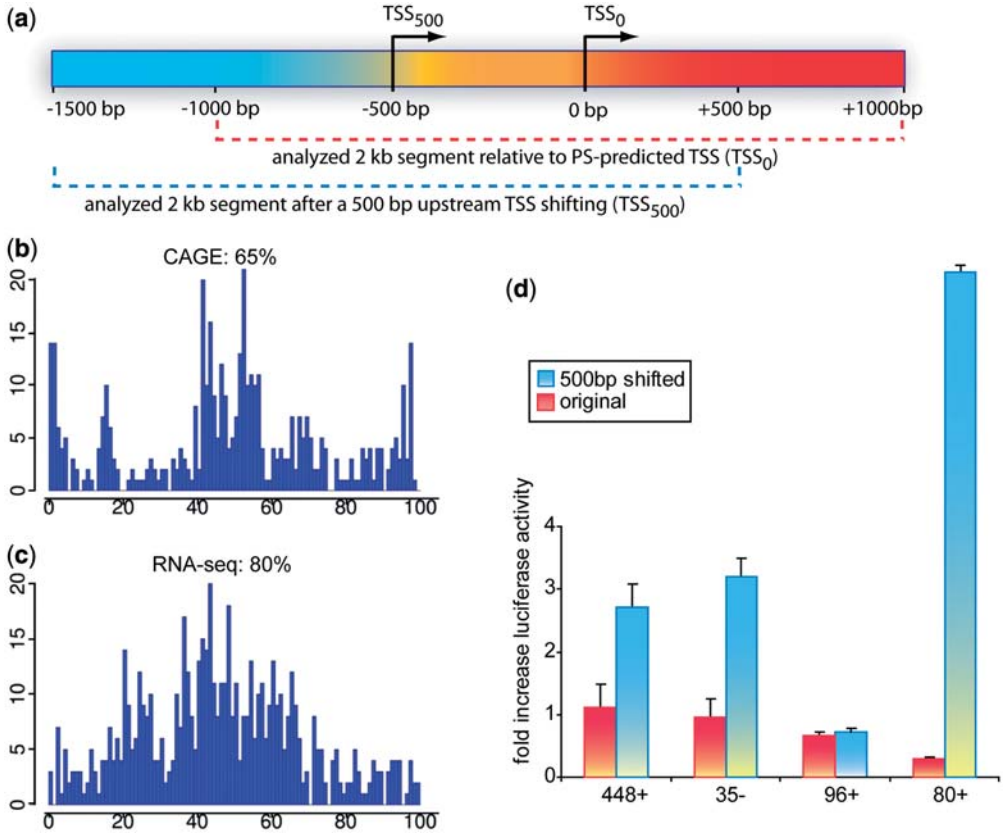


Figure 4. Evaluation of PS+L− sequences on centering the TSS 500 bp upstream from the prediction. (**a**) Subset 2–shifted regions were reconstructed by first re-locating the TSS 500 bp upstream from the relative prediction in the human genome, and subsequently selecting the flanking ±1000 bp upstream and downstream regions, respectively (**b**) Distribution of CAGE tags in H1-hESC cells for the 2000 bp regions centered in relocated TSSs (**c**) RNA-seq analysis profiles of the same regions. X-axes show % distance bins, each one including 20 bp. Y-axes display the number of detected tags. Here we observe a major peak from both analyses around the 50th bin (1000 bp), indicating that it may correspond to a transcription start region (**d**) We confirmed the transcription ability of those regions by additional luciferase assays in four representative PS+L− sequences (three in sense strand and one in anti-sense), showing a significant higher activity (green bars) as compared with the original predictions (red).

Core promoter activity landscape

We subsequently analyzed the four subsets of ProStar predictions to seek for correlations between structural or epigenetic motifs and the promoter activity status. To this end, we used data repositories publicly available from the ENCODE Consortium (9) (See 'Materials and Methods' section for details).

We first analyzed chromatin accessibility to DNase I degradation profiles, as DNase I hypersensitive sites (DHS) are expected to correlate with loosely packed regions in bulk chromatin and hence with gene transcriptional activity (34–36). Analysis of ENCODE data (Figure 5a) highlighted a similar DHS density for ProStar positive subsets (PS+L+ and PS+L−), which turned out to be much larger than the observed density for the negative subsets (PS−L+ and PS−L−). These observations indicate that ProStar-predicted regions are indeed open and thereby associated with transcriptionally active chromatin. Of note, those predictions cannot be simply explained on the basis of sequence-dependent rules such as the presence of CpG islands, as the CG content provides a disperse prediction signal and leads to large number of false positives (13). It should also be noted that ProStar is able to detect promoters located at a large distance to any annotated CpG island (37), as this is the case for 20% of the positive predictions analyzed by CAGE (Supplementary Table S2).

We also evaluated the occurrence of histone modifications correlated with epigenetic modulation of gene transcription, in particular H3K4Me1, H3K27Ac and H3K4Me3, which are specifically prevalent in regulatory regions (38). The results shown in Figure 5b revealed that these histone marks were actually more overrepresented in ProStar positive regions (PS+L+ and PS+L−) than in ProStar negative regions (PS−L+ and PS−L−), providing accumulating evidence about the attainable implication of ProStar regions in the regulation of gene activity.

Furthermore, as PS+ regions are located on regulatory elements, we quested for potential associations to specific functions by a TFBS enrichment evaluation, using the Transfac database (27). To this end, we examined diverse region subsets, including all PS high-scored predictions (PS+, 17909 sequences), the experimentally tested

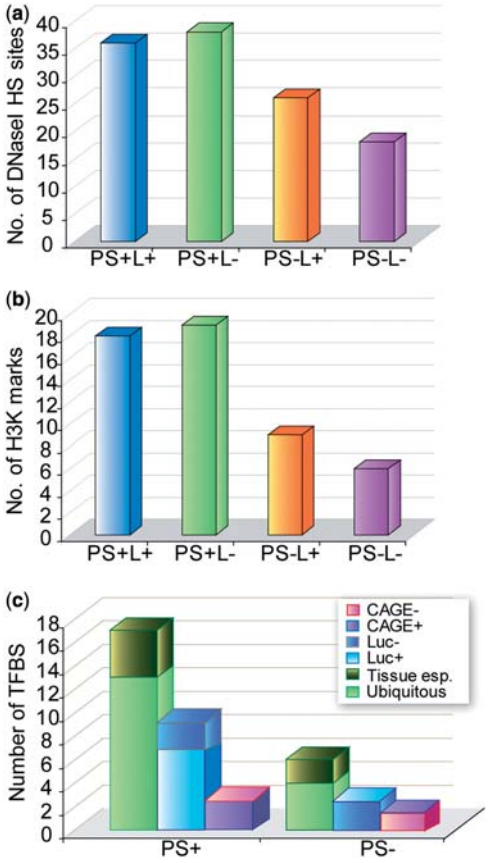


Figure 5. Putative promoter activity landscape. (**a**) Average DHS enrichment within 1200 bp regions of the different CAGE analyzed subsets in a large collection of cell types available in ENCODE (**b**) Average plots of detected Histone 3 variants correlating with transcriptional activity: H3K4Me1, often observed near regulatory elements; H3K27Ac, occurring near promoters; H3K4Me3, near active regulatory elements (**c**) TFBS enrichment evaluation of PS+ and PS− predictions according to gene expression (i.e. ubiquitous vs tissue-specific; left column in the respective PS+ and PS− groups). Further evaluation of TFBS enrichment in the experimentally tested PS+ and PS− sequence sets according to luciferase transcription activity (Luc+ or Luc−, middle bars) and CAGE mapping analysis (CAGE+ or CAGE−, right bars).

regions (containing 119 PS+ predictions and 100 PS− predictions, respectively) and the CAGE-analyzed sets (subsets 1 and 2 on the one hand, and subsets 3 and 4 on the other hand). The enrichment for a given TFBS was considered to be significant when $P < 5.65 \times 10^{-5}$ (Supplementary Table S3). Again, PS− regions showed little enrichment, whereas 17 human TFBSs were found to be overrepresented in at least one of the PS+ groups, the larger part being annotated as ubiquitous (Figure 5c, left column in PS+ and PS− groups, respectively) and mostly related to vital cellular functions (Supplementary Table S3). Interestingly, TFBSs overrepresented in the regions capable of driving luciferase transcription were also enriched in PS+ predictions (Figure 5c, middle bars, Luc+) and CAGE tagged sequences (Figure 5c, right bars, CAGE+). In addition, the identified TFBSs presented high binding affinity to GC-rich sequences, representing truly active TFBSs and thereby supporting our hypothesis that ProStar accurately predicts promoters of housekeeping genes.

Lastly, although the vast majority of the PS+ regions were not detectable using phylogenetic footprinting-based methods (as we previously discussed), we further investigated the conservation of ProStar regions across species, as biologically relevant sequences should display some level of sequence conservation. As expected, PS+ regions were enriched for conserved DNA elements (Figure 6a), particularly for TFBSs (Figure 6b), as compared with PS− regions.

DISCUSSION

A comprehensive analysis of ProStar predicted TSSs has enabled us to identify novel functional core promoters in the human genome exclusively detected by their differential physical deformability pattern and not simply by sequence-based signals such as the CG content alone. A large percentage of ProStar seemingly 'false positives', i.e. regions with unusual physical properties but not associated to any annotated promoter, are indeed transcriptionally active. In particular, highly active regions containing a differential physical pattern display typical chromatin features of housekeeping gene promoters involved in cell survival and maintenance, as proven by an overwhelming amount of direct (luciferase assays, CAGE or RNA-seq mapping) and indirect evidence (profile analysis such as DNaseI sensitivity, epigenetic markers, TFBS enrichment or DNA element conservation). Interestingly, physical signaling also appears to be able to detect promoter activity even in cases where the TSS is located 500 bp upstream of the prediction. Whether this displacement is indicative of a particular feature of genes with closely related alternative TSSs, as indicated by massive CAGE and RNA-seq mappings (Figures 3b and 4), will nevertheless require further investigation. Taken together, these observations reinforce the evidence that high-confidence ProStar predicted regions, sharing a defined pattern of physical features, truly behave like physiologically active TSSs.

We have also observed that most of the active core regions signaled by physical properties do not exhibit directionality in transcript initiation, indicating that physical properties might signal zones where the binding of regulatory proteins and the deformation of DNA are less intricate, as we had previously suggested (39–42). Yet, this signaling might not be sufficient to determine the correct sense of transcription. Intriguingly, more than half of all human promoters are bidirectional, and hence directionality of promoter activity may be regulated to some degree in a cell type–specific manner (43).

On the whole, our study provides insights into the role of DNA physical properties in ascertaining an ancestral coarse regulatory mechanism. Thereby, regions with high chance of undergoing spontaneous transcription would be recognized by protein effectors and favor nucleosome depletion aside from the purely sequence-based signals encoded as H-bond patterns in the DNA major and

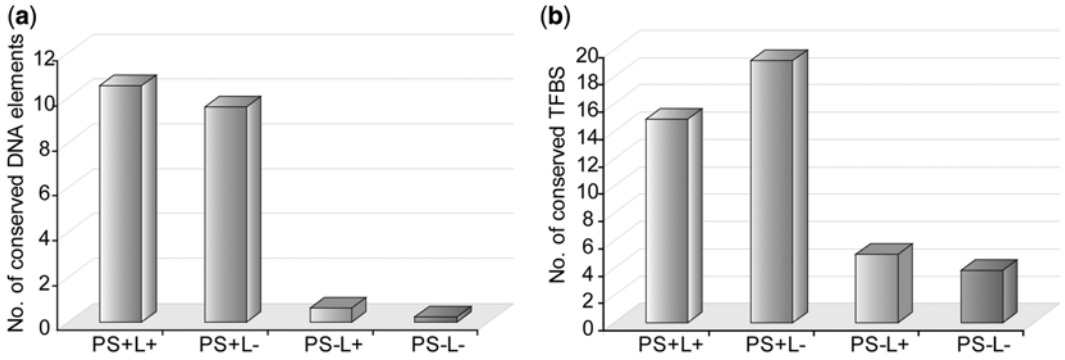


Figure 6. Putative promoter interspecies conservation. (**a**) Average plots of identified DNA elements conserved among human, mouse and rat using the 'Vetebrate PhyloP' algorithm within 1200 bp regions of the different CAGE analyzed subsets (**b**) Similarly, plot histograms showing the average number of identified TFBSs that are conserved across species.

minor grooves. In fact, recent genome-wide nucleosome mapping analyses from our group have revealed that housekeeping genes display unique nucleosome architectures, with large nucleosome refractory regions upstream the TSS (unpublished data). In general then, the interplay between DNA physical properties and regulatory regions could be rationalized in terms of nucleosome positioning (16), favoring the presence of sequences with unique deformation properties in promoter regions, although this might not be the only underlying mechanism, and this would probably vary from gene to gene.

Yet, the physical code type of mechanism could have been evolutionary deactivated in specific genes where fine regulation is required, but seems to be still active in many other cases, where such a stringent regulation is not essential. This convoluted regulatory signaling present in complex organisms could partially explain the failure of traditional promoter location methods to identify a significant number of TSSs, implying the presence of many hidden promoter regions in the human genome.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–3 and Supplementary Figures 1 and 2.

ACKNOWLEDGEMENTS

We thank Carles Fenollosa and Ramon Goñi for technical support with ProStar.

FUNDING

Spanish Ministry of Science and Innovation [BIO2012-32868 and Consolider E-Science Project]; Instituto de Salud Carlos III (Instituto Nacional de Bioinformática); European Research Council (ERC) Advanced Grant; Fundación Marcelino Botín. M.O. is an Institució Catalana de Recerca i Estudis Avançats (ICREA) Academia Researcher. Funding for open access charge: Fundación Marcelino Botín.

Conflict of interest statement. None declared.

REFERENCES

1. Hélène,C. (1981) Recognition of base sequences by regulatory proteins in procaryotes and eucaryotes. *Biosci. Rep.*, **1**, 477–483.
2. Baumann,M., Pontiller,J. and Ernst,W. (2010) Structure and basal transcription complex of RNA polymerase II core promoters in the mammalian genome: an overview. *Mol. Biotechnol.*, **45**, 241–247.
3. Hannenhalli,S. and Levy,S. (2001) Promoter prediction in the human genome. *Bioinformatics*, **17**, S90–S96.
4. Davuluri,R.V., Grosse,I. and Zhang,M.Q. (2001) Computational identification of promoters and first exons in the human genome. *Nat. Genet.*, **29**, 412–417.
5. Bajic,V.B., Brent,M.R., Brown,R.H., Frankish,A., Harrow,J., Ohler,U., Solovyev,V.V. and Tan,S.L. (2006) Performance assessment of promoter predictions on ENCODE regions in the EGASP experiment. *Genome Biol.*, **7**, S3.
6. Bajic,V.B., Tan,S.L., Suzuki,Y. and Sugano,S. (2004) Promoter prediction analysis on the whole human genome. *Nat. Biotechnol.*, **22**, 1467–1473.
7. Cooper,S.J., Trinklein,N.D., Anton,E.D., Nguyen,L. and Myers,R.M. (2006) Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Res.*, **16**, 1–10.
8. Birney,E., Stamatoyannopoulos,J.A., Dutta,A., Guigó,R., Gingeras,T.R., Margulies,E.H., Weng,Z., Snyder,M., Dermitzakis,E.T. and Thurman,R.E. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
9. Raney,B.J., Cline,M.S., Rosenbloom,K.R., Dreszer,T.R., Learned,K., Barber,G.P., Meyer,L.R., Sloan,C.A., Malladi,V.S. and Roskin,K.M. (2011) ENCODE whole-genome data in the UCSC genome browser (2011 update). *Nucleic Acids Res.*, **39**, D871–D875.
10. Schueler,M.G. and Sullivan,B.A. (2006) Structural and functional dynamics of human centromeric chromatin. *Annu. Rev. Genomics Hum. Genet.*, **7**, 301–313.
11. Butcher,L.M. and Beck,S. (2008) Future impact of integrated high-throughput methylome analyses on human health and disease. *J. Genet. Genomics*, **35**, 391–401.
12. Pedersen,A.G., Baldi,P., Chauvin,Y. and Brunak,S. (1998) DNA structure in human RNA polymerase II promoters1. *J. Mol. Biol.*, **281**, 663–673.
13. Goñi,J.R., Pérez,A., Torrents,D. and Orozco,M. (2007) Determining promoter location based on DNA structure first-principles calculations. *Genome Biol.*, **8**, R263.
14. Abeel,T., Saeys,Y., Bonnet,E., Rouzé,P. and Van de Peer,Y. (2008) Generic eukaryotic core promoter prediction using structural features of DNA. *Genome Res.*, **18**, 310–323.
15. Goñi,J.R., Fenollosa,C., Pérez,A., Torrents,D. and Orozco,M. (2008) DNAlive: a tool for the physical analysis of DNA at the genomic scale. *Bioinformatics*, **24**, 1731–1732.

16. Deniz,O., 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.
17. Boeger,H., Griesenbeck,J., Strattan,J.S. and Kornberg,R.D. (2003) Nucleosomes unfold completely at a transcriptionally active promoter. *Mol. Cell*, **11**, 1587–1598.
18. Gross,P. and Oelgeschläger,T. (2006) Core promoter-selective RNA polymerase II transcription. *Biochem. Soc. Symp.*, **73**, 225–236.
19. Harrow,J., Denoeud,F., Frankish,A., Reymond,A., Chen,C.K., Chrast,J., Lagarde,J., Gilbert,J.G., Storey,R. and Swarbreck,D. (2006) GENCODE: producing a reference annotation for ENCODE. *Genome Biol.*, **7**, S4.
20. Marques de Sa,J.P. (2001) *Pattern Recognition: Concepts, Methods, and Applications*. Springer Verlag, Berlin.
21. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V. and Down,T. (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.
22. Mancuso,G. and Rugarli,E. (2008) A cryptic promoter in the first exon of the SPG4 gene directs the synthesis of the 60-kDa spastin isoform. *BMC Biol.*, **6**, 31.
23. Consortium,E.P. (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biology*, **9**, e1001046.
24. Djebali,S., Lagarde,J., Kapranov,P., Lacroix,V., Borel,C., Mudge,J.M., Howald,C., Foissac,S., Ucla,C. and Chrast,J. (2012) Evidence for transcript networks composed of chimeric RNAs in human cells. *PLoS One*, **7**, e28213.
25. Djebali,S., Davis,C.A., Merkel,A., Dobin,A., Lassmann,T., Mortazavi,A.M., Tanzer,A., Lagarde,J., Lin,W., Schlesinger,F. *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.
26. Karolchik,D., Hinrichs,A.S., Furey,T.S., Roskin,K.M., Sugnet,C.W., Haussler,D. and Kent,W.J. (2004) The UCSC table browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493–D496.
27. Wingender,E., Dietze,P., Karas,H. and Knüppel,R. (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, **24**, 238–241.
28. González,S., Montserrat-Sentís,B., Sánchez,F., Puiggròs,M., Blanco,E., Ramirez,A. and Torrents,D. (2012) ReLA, a local alignment search tool for the identification of distal and proximal gene regulatory regions and their conserved transcription factor binding sites. *Bioinformatics*, **28**, 763–770.
29. Daniel,F.I., Cherubini,K., Yurgel,L.S., de Figueiredo,M.A.Z. and Salum,F.G. (2011) The role of epigenetic transcription repression and DNA methyltransferases in cancer. *Cancer*, **117**, 677–687.
30. Hawkins,P. and Morris,K.V. (2008) RNA and transcriptional modulation of gene expression. *Cell Cycle*, **7**, 602.
31. Fukuda,H., Sano,N., Muto,S. and Horikoshi,M. (2006) Simple histone acetylation plays a complex role in the regulation of gene expression. *Brief. Funct. Genomics Proteomics*, **5**, 190–208.
32. Kodzius,R., Kojima,M., Nishiyori,H., Nakamura,M., Fukuda,S., Tagami,M., Sasaki,D., Imamura,K., Kai,C. and Harbers,M. (2006) CAGE: cap analysis of gene expression. *Nat. Methods*, **3**, 211–222.
33. Trinklein,N.D., Karaöz,U., Wu,J., Halees,A., Aldred,S.F., Collins,P.J., Zheng,D., Zhang,Z.D., Gerstein,M.B. and Snyder,M. (2007) Integrated analysis of experimental data sets reveals many novel promoters in 1% of the human genome. *Genome Res.*, **17**, 720–731.
34. Sabo,P.J., Humbert,R., Hawrylycz,M., Wallace,J.C., Dorschner,M.O., McArthur,M. and Stamatoyannopoulos,J.A. (2004) Genome-wide identification of DNaseI hypersensitive sites using active chromatin sequence libraries. *Proc. Natl Acad. Sci. USA*, **101**, 4537.
35. Boyle,A.P., Davis,S., Shulha,H.P., Meltzer,P., Margulies,E.H., Weng,Z., Furey,T.S. and Crawford,G.E. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311–322.
36. Dans,P.D., Pérez,A., Faustino,I., Lavery,R. and Orozco,M. (2012) Exploring polymorphisms in B-DNA helical conformations. *Nucleic Acids Res.*, **40**, 10668–10678.
37. Meyer,L.R., Zweig,A.S., Hinrichs,A.S., Karolchik,D., Kuhn,R.M., Wong,M., Sloan,C.A., Rosenbloom,K.R., Roe,G. and Rhead,B. (2013) The UCSC genome browser database: extensions and updates 2013. *Nucleic Acids Res.*, **41**, D64–D69.
38. Rice,J.C. and Allis,C.D. (2001) Histone methylation versus histone acetylation: new insights into epigenetic regulation. *Cur. Opin. Cell. Biol.*, **13**, 263–273.
39. 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**, 63.
40. 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**, 627–638.
41. 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**, 6144–6151.
42. Orozco,M., Noy,A. and Pérez,A. (2008) Recent advances in the study of nucleic acid flexibility by molecular dynamics. *Cur. Opin. Struct. Biol.*, **18**, 185–193.
43. Trinklein,N.D., Aldred,S.F., Hartman,S.J., Schroeder,D.I., Otillar,R.P. and Myers,R.M. (2004) An abundance of bidirectional promoters in the human genome. *Genome Res.*, **14**, 62–66.

2. Nucleosome organization in vivo

In the previous section, we presented how the theoretical modeling of DNA was validated with experimental results. This implies that those experiments were designed after a hypothesis developed in the computational lab which was later validated in the wet lab.

Now, we will present some works born in the wet lab but which required a large bioinformatics support in their analysis. In this section we will review two works done in equally contribution with Özgen Deniz from EBL, as well as two collaborations with external groups.

In the first article of this section, *Nucleosome architecture and plasticity along cell cycle* (page 74) our objective was to obtain nucleosome maps together with gene expression information in the different stages of the cell. This will provide us with an accurate atlas of the dynamics of chromatin during the processes of DNA replication and chromosome segregation, which could be strong determinants of the nucleosome structure.

After few years working with nucleosomes, we realized that a periodic problem in our research was to differentiate negligible intrinsic noise effects from the significant biological variations in different experiments. As we accounted with different replicas of cell-cycle synchronized nucleosome maps, additionally to other asynchronous maps, we systematically analyzed and quantified the variability observed in the maps obtained under different technical and biological conditions. The conclusions are presented in the work *Fuzziness and noise in nucleosomal architecture* (page 78).

Finally, we will present two collaborations with external groups, featuring relevant conclusions regarding chromatin organization. The first one is a publication done in collaboration with Prof. Francesc Posas of the Pompeu Fabra University in Barcelona titled *Hog1 bypasses stress-mediated down-regulation of transcription by RNA polymerase II redistribution and chromatin remodeling* (page 92). For the second one I will present some of the results obtained in the collaboration performed during a short stay in the group of Prof. Jason Lieb from the University of North Carolina in Chapel Hill, USA. This can be found in the section entitled *Modeling genome-wide kinetics of endo/exonuclease digestion in C. elegans* (page 105).

2.1. Nucleosome architecture and plasticity along cell cycle

Proliferation of all cells is mediated through the cell-division cycle, which consists of four main phases: genome duplication (S phase) and nuclear division (mitosis or M phase), separated by two gap phases (G1 and G2) (Fig. 21).

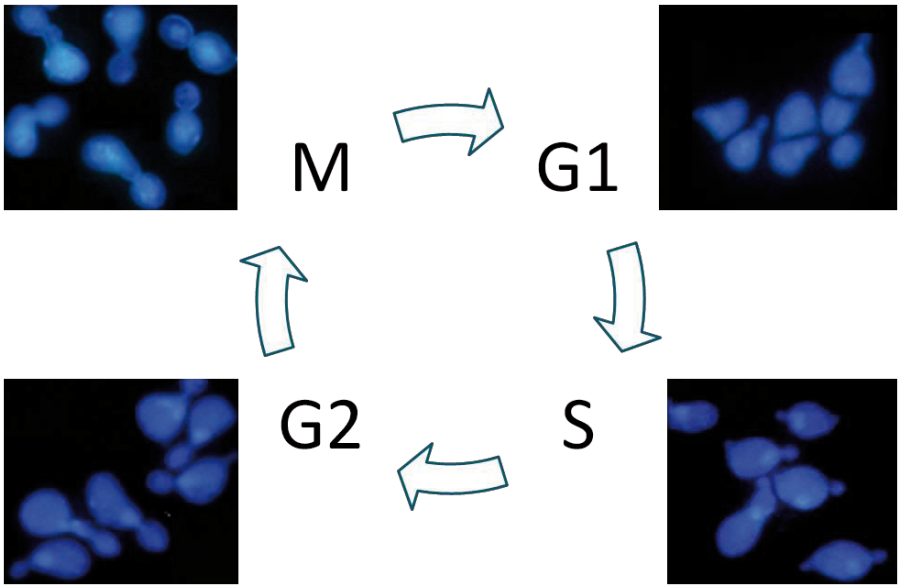


Fig. 21 Yeast cells budding during cell cycle. Images obtained with fluorescent microscopy over synchronized cell population.

Although cell cycle–regulated transcription has been worked out in greater detail for yeast, its integration with chromatin structure remains uncertain (Bähler, 2005). Pioneering work by Hogan and coworkers (Hogan *et al.*, 2006) was very insightful since it first established nucleosome occupancy fluctuations around promoters throughout the cell cycle. However, this early approach was only able to resolve NFRs but not individual nucleosomes and their analysis was limited to a few genes.

The motivation to study chromatin dynamics in a population of cell-cycle synchronized cells is interesting for different reasons: 1) chromatin suffers drastic changes during the synthesis (S) and mitosis (M) stage of the cycle, when all DNA is replicated and segregated; 2) as said before, periodic genes involved in cell cycle are well studied and their patterns of expression can be linked with the dynamic changes in the chromatin and 3) working with a synchronized population of cells decrease the noise in the coverage profile due to the factors previously mentioned.

In this work we synchronized a population of yeast cells arresting them in G1 phase by adding alpha-factor mating pheromone. After the alpha-factor release, samples were collected at 0, 30, 40 and 50 minutes and correct cell-cycle synchronization was validated by Fluorescence-activated Cell Sorting (FACS) and microscopy inspection. Nucleosomal DNA and mRNA were extracted from the different samples and sent to the sequencing facility, allowing us to obtain nucleosome maps and gene expression levels from different stages of the cell-cycle. This enabled the first high-resolution study of nucleosome position and plasticity throughout the cell cycle in budding yeast.

Despite this work was started some time ago, experimental problems with the biological replicates delayed the redaction of the final manuscript and motivated us to publish some of the observed results in a different paper. This parallel work focused in the technical and biological factors which affect the general fuzziness of nucleosome maps (see the work *Fuzziness and noise in nucleosomal architecture* on page 78). At the moment of the deposit of this thesis, the manuscript of this work is in preparation.

Our observations point that changes in chromatin during the cell cycle are not massive and they are focused in certain genes. In global trends, G1 stage shows a better nucleosome positioning meanwhile S and M stages have a significantly fuzzier profile (Fig. 22). The relationship between the dynamics of gene expression and the chromatin structure are not so evident and require a detailed gene by gene analysis. One of the interesting trends we found regarding this point is that genes with higher basal levels of expression show a relatively higher coverage in S and M stages compared to G1 and G2.

Looking at chromatin dynamics between consecutive stages, we found that genes with high changes in the nucleosome coverage in the TSS between G1 and S stages are significantly enriched in genes linked to stress-response pheromones, probably caused by the use of alpha-factor in the arrest of cells. More examples of the link with chromatin dynamics and gene function are the higher plasticity of genes changing its nucleosomal structure between G2 and M, enriched in GO terms associated to cell-wall biogenesis.

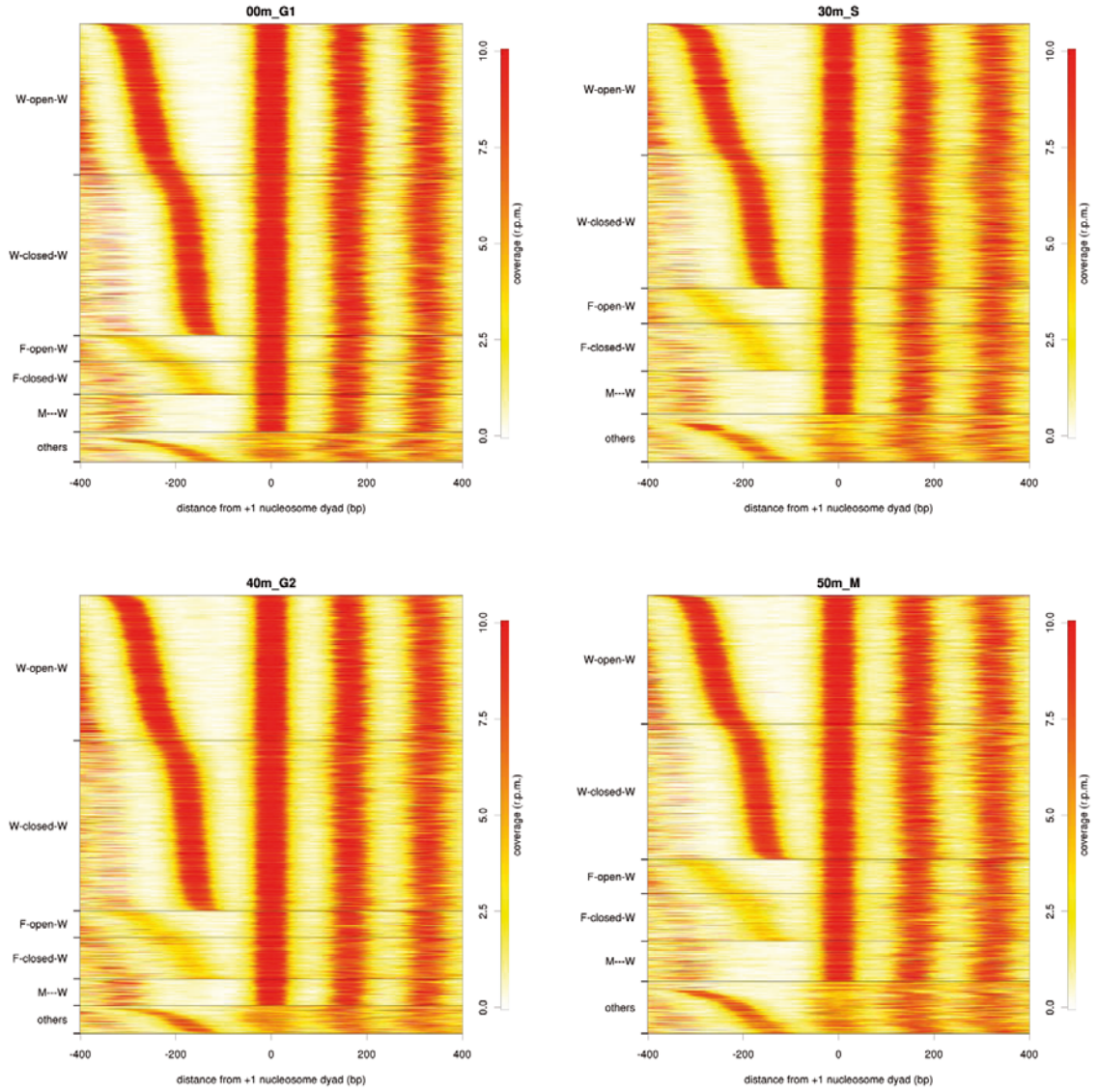


Fig. 22 Genome-wide nucleosome profile classification around +1 nucleosome of G1 (top-left), S (top-right), G2(bottom-left) and M (bottom-right) stages. Fuziness is increased in S and M samples, while G1 and G2 display a better positioning.

Genes with known periodic expression along cell cycle are also enriched in dynamic chromatin structures. After the alpha-factor release, cyclins CLN1 and CLN2 block the SIC1 inhibitor which suppresses CDK1 activity in S stage. After G2 stage, the mRNA levels of SIC1 increase and this change is coupled to a nucleosome eviction around the TSS of that gene (Fig. 23). The block of SIC1 promotes the transcription of CLB5 and CLB6, active during the S stage and increased again in M stage (Fig. 23). CLB5 and CLB6 promote the DNA replication by triggering the pre-replicative com-

plex in replication origins, which, after its activation, requires the activity of helicase maintenance (MCM) proteins to expose template DNA. Changes in the nucleosomal architecture coupled to changes in the expression levels can be accordingly observed in MCM3 and MCM7 genes in G1/M stages (Fig. 23).

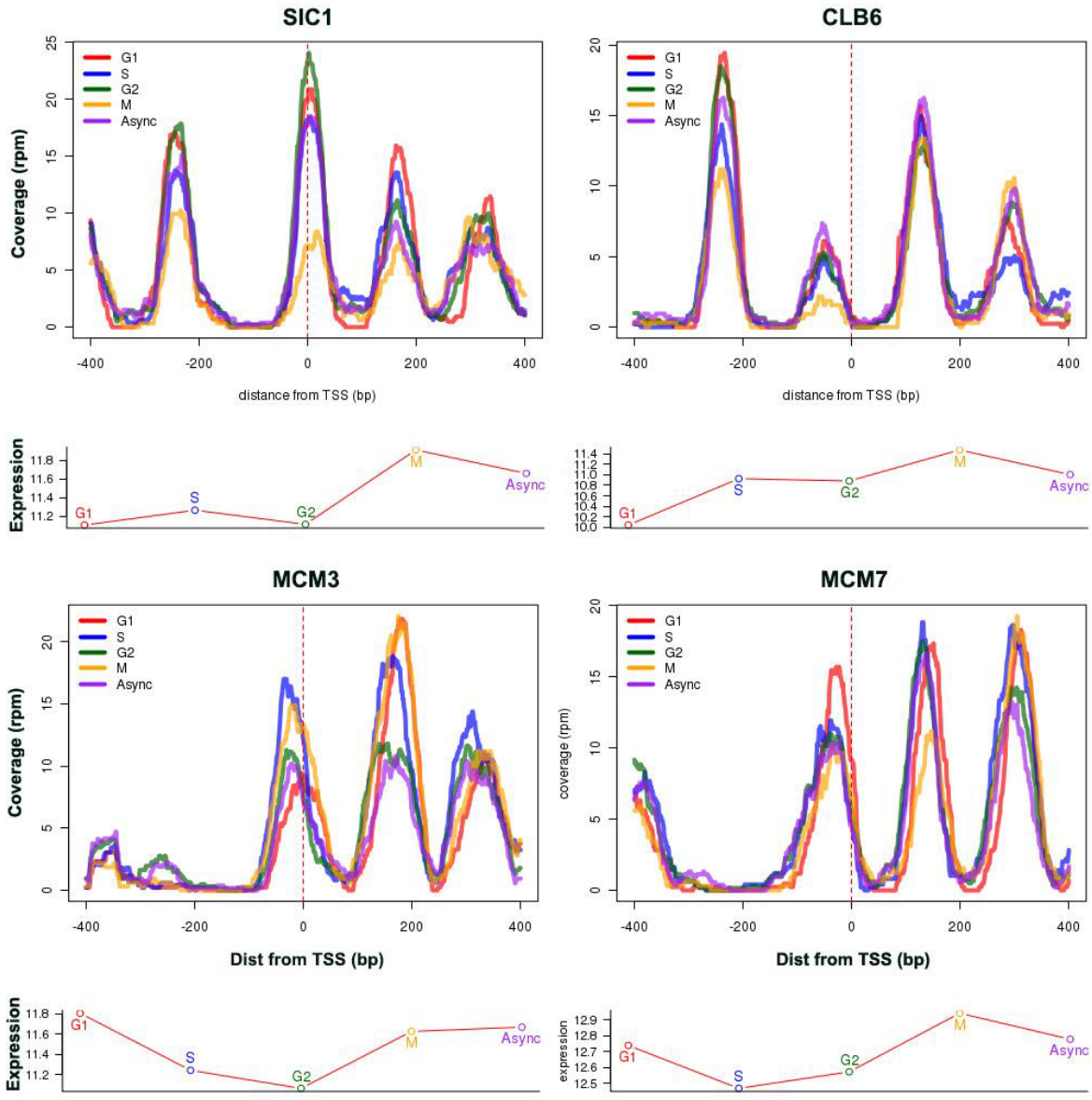


Fig. 23. Individual nucleosome occupancy and gene expression levels for genes SIC1, CLB6, MCM3 and MCM7 along cell cycle

Publication:

Deniz, Ö.*, Flores, O.*, Soler-López, M., Orozco, M. "*Nucleosome architecture and plasticity along cell cycle*", (in preparation).

2.2. Fuzziness and noise in nucleosomal architecture

From 2009 to 2014 we processed and analyzed more than 56 different nucleosome maps done in our group and around 40 more obtained from other groups, either as collaborations or downloaded from on-line repositories. This motivated the development of methods and pipelines for the automatic processing and bulk analysis of such data, but also allowed us to gain expertise in dealing with different technical and biological problems linked to these experiments.

As seen previously (Fig. 12, page 18), extracting the DNA from a population of cells forces the study of an averaged positioning of the nucleosomes, yielding to the concept of nucleosome fuzziness. Of course, this fuzziness can be caused either by technical noise or due to intrinsic biologic factors as the chromatin remodelers or chromosome dynamics during the cell cycle.

In order to define and quantify methodologically those sources of variability inside and between nucleosome maps, in this work, we systemically analyzed 7 nucleosome maps with controlled changes in certain conditions to measure the effect of different sources of noise. Additionally, the application of DNA mechanics introduced in the previous chapter revealed an underlying mechanism for determination of the positioning of the nucleosomes.

The studied sources of noise include (Fig. 24):

- Reads mapping in opposite strands in single-end sequencing. Despite strand correction works fine in a normal case, shorter or longer reads will be misaligned causing a fuzzier coverage peak (Fig. 24a).
- The presence of long fragments in paired-end sequencing. If reads are not filtered according to their width, we could capture di-nucleosome fragments with their dyad mapping in the linker region between two nucleosomes. Of course, this is only a problem when reads are trimmed, but this is a usual process in to increase the signal-to-noise ratio (Fig. 24b).
- Related to the last two points, differential chromatin digestion with Micrococcal Nuclease would have different effects on the observed nucleosome coverage.
- As presented in the previous work, inherent chromatin dynamics linked to the cell cycle can be an additional source of noise respect synchronized cell populations (Fig. 24d).

- Finally, our observations regarding the intrinsic flexibility of DNA revealed energetic barriers which could constrain the intrinsic nucleosome sliding and, therefore, allow a better phasing of the neighboring nucleosomes (Fig. 24c).

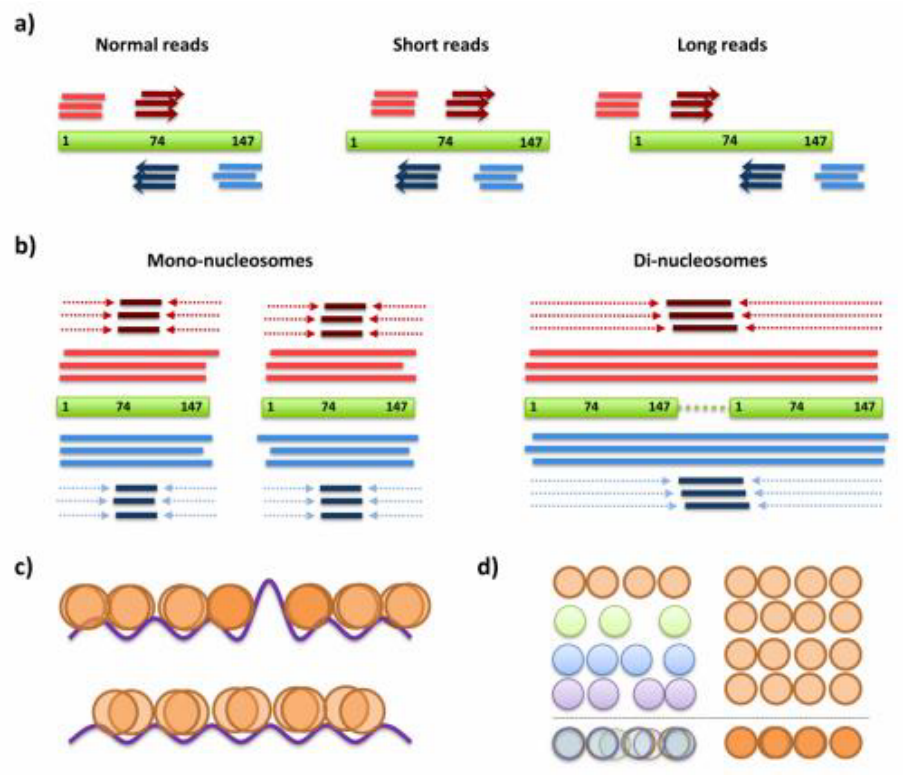


Fig. 24. Sources of fuzziness and noise in the nucleosome maps. a) Single-end sequencing misalignment b) Presence of di-nucleosomes in paired-end sequencing, c) energetic barriers act as a positioning element and d) cells with heterogeneous growth stages add an additional source of noise compared to synchronous populations.

Noteworthy, a clustering method based on the classification of +1/-1 nucleosomes (Fuzzy, F, or Well-positioned, W) together with a measure of the distance between them (open or closed) (Zaugg and Luscombe, 2011), combined with the mechanistic model presented in the section *Experimental impact of theoretical DNA mechanics* (page **Error! Bookmark not defined.**), revealed two canonical conserved structures in promoters with different effect on the nucleosome phasing. *Open* promoters feature two clear energy barriers which strongly position the -1 and +1 nucleosomes. Then, we can reproduce the positioning of upstream and downstream nucleosomes with a simple linear statistical model (Fig. 25a). In the other hand, *Closed* promoters present a different mechanism which cannot be fitted to a simple model. In both cases, the nucleosome positioning can be

predicted with a combination of intrinsic physical properties combined with information of putative transcription factor binding (Fig. 25b).

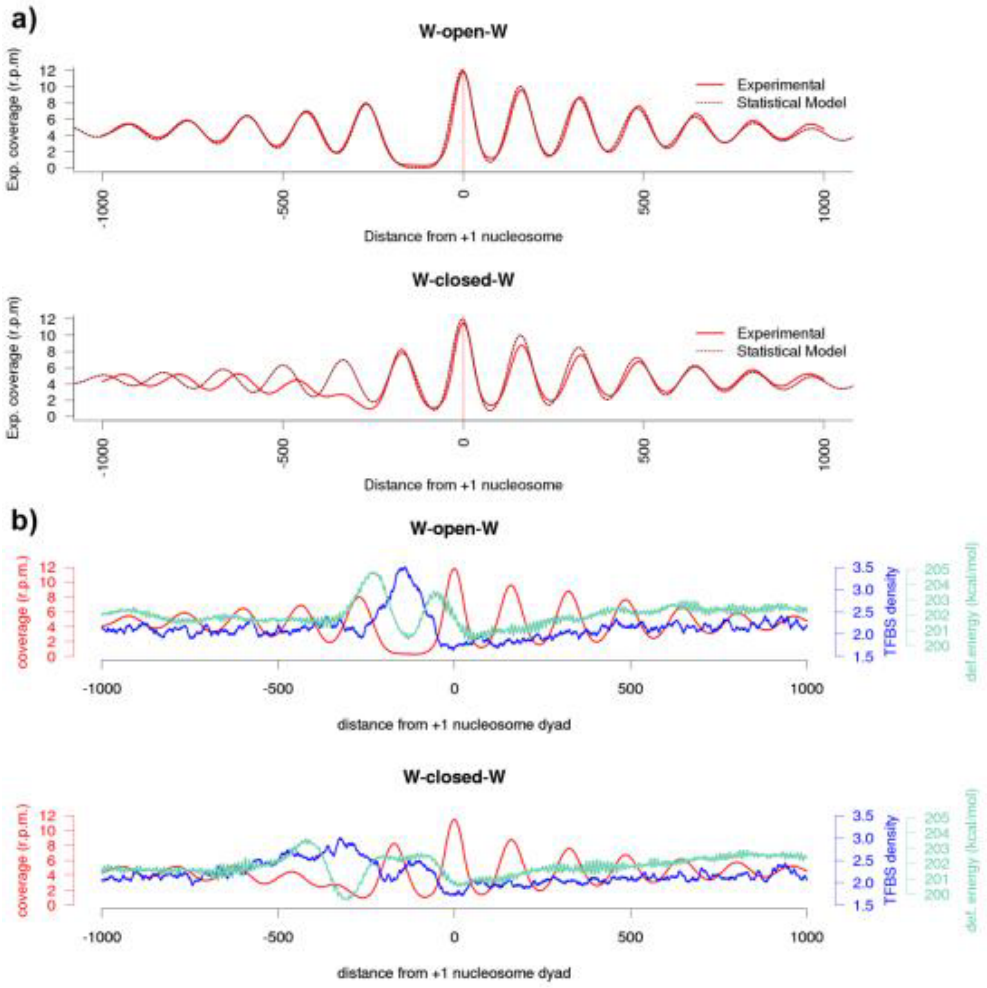


Fig. 25. a) Fitting of a simple statistical model of positioning in open and closed promoter configurations. b) The positioning of the -1 and +1 nucleosomes is explained with a cooperative effect of the DNA intrinsic deformation energy and putative Transcription Factor Binding Sites (TFBS), which both repel the nucleosomes acting as a phasing element.

Publication:

Flores, O.*, Deniz, Ö*, D. Soler-López, M., Orozco, M. (2014) "*Fuzziness and noise in nucleosomal architecture*", Nucleic Acids Research (early access).

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

Nucleic Acids Research Advance Access published February 27, 2014

Nucleic Acids Research, 2014, 1–13
doi:10.1093/nar/gku165

Fuzziness and noise in nucleosomal architecture

Oscar Flores[1,2,†], Özgen Deniz[1,2,†], Montserrat Soler-López[1,2] and Modesto Orozco[1,2,3,*]

[1]Institute for Research in Biomedicine (IRB Barcelona), Baldiri Reixac 10-12, 08028 Barcelona, Spain, [2]Joint IRB-BSC Program in Computational Biology, Baldiri Reixac 10-12, 08028 Barcelona, Spain and [3]Department of Biochemistry and Molecular Biology. University of Barcelona, 08028 Barcelona, Spain

Received November 12, 2013; Revised February 7, 2014; Accepted February 8, 2014

ABSTRACT

Nucleosome organization plays a key role in the regulation of gene expression. However, despite the striking advances in the accuracy of nucleosome maps, there are still severe discrepancies on individual nucleosome positioning and how this influences gene regulation. The variability among nucleosome maps, which precludes the fine analysis of nucleosome positioning, might emerge from diverse sources. We have carefully inspected the extrinsic factors that may induce diversity by the comparison of micrococcal nuclease (MNase)-Seq derived nucleosome maps generated under distinct conditions. Furthermore, we have also explored the variation originated from intrinsic nucleosome dynamics by generating additional maps derived from cell cycle synchronized and asynchronous yeast cultures. Taken together, our study has enabled us to measure the effect of noise in nucleosome occupancy and positioning and provides insights into the underlying determinants. Furthermore, we present a systematic approach that may guide the standardization of MNase-Seq experiments in order to generate reproducible genome-wide nucleosome patterns.

INTRODUCTION

Eukaryotic chromatin is organized in a compact, precisely regulated, but yet not fully understood manner. Nucleosome is the fundamental structural unit of this compaction (1,2), formed by the wrapping of 147-bp double-stranded DNA in 1 and ¾ left-handed superhelix around a histone octamer (3). The presence of histone proteins determines the accessibility of DNA to other interacting proteins and plays a role in altering mutation rate (4), in determining exon architecture (5) and in regulation of transcription (6,7).

Genome-wide studies of micrococcal nuclease (MNase) susceptibility have revealed that nucleosomes are not randomly placed across genome but they are rather enriched in certain areas while depleted in others (8–13). The most pervasive regions depleted in nucleosomes appear upstream the transcription start site (TSS), at gene promoters. Such nucleosome-free regions (NFRs) are often flanked by two nucleosomes, one very strongly located downstream TSS (+1 position), and a second one more weakly located upstream (−1 position) (14,15,8,10,13). The integrity of this organization seems to be crucial for a correct gene regulation (16–18), since the introduction of a high-affinity nucleosome binding sequence in a promoter region can inhibit transcription (19).

Despite the tremendous amount of studies, the underlying mechanisms governing *in vivo* nucleosome positioning still remain elusive. Three main models have been postulated as determinants of nucleosome positioning: (i) statistical positioning, suggesting that a barrier favors deposition of a well-positioned nucleosome, typically after a NFRs, which in turn forces the periodic positioning of the neighboring nucleosomes (20,21); (ii) the intrinsic properties of naked DNA that favors histone binding in certain sequences (8,12,22–28) and (iii) DNA-interacting proteins like transcription factors (TFs) (11,29–31), which force nucleosome depletion in certain regions. Notably, there is still controversy among these three models, since all of them seem supported by experimental findings (32–36).

Systematic analyses of nucleosome dynamics still remain a challenge, particularly in determining the exact positioning, fuzziness and affinity (37–42). The struggle partially arises from the paucity of consistent nucleosome maps, even in the same model organisms. Thus, different studies show quite similar average nucleosome profiles, but individual nucleosome positioning might differ remarkably, so that very well-positioned nucleosomes detected in one study might appear as fuzzy or simply absent in another (43–46). As an example, a recent study showed that Pearson's correlation coefficients range from 0.2 to 0.45 among different *in vivo* nucleosome profile

*To whom correspondence should be addressed. Tel: +34 934 037 156; Fax: +34 934 037 157; Email: modesto.orozco@irbbarcelona.org

†These authors contributed equally to the paper as first authors.

datasets, indicating that in average <10% of nucleosome positioning is actually reproduced by different studies (47). The extreme variability among datasets, not only makes difficult the derivation of consensus nucleosome maps, but also it just makes impossible the development of predictive models to explain nucleosome positioning.

Discrepancy between nucleosome maps may originate from different sources: (i) the experimental conditions (such as MNase digestion levels or sequencing protocol); (ii) data processing for nucleosome calling; (iii) heterogeneity of the samples, derived from diversity of cellular states in the culture, such as cycle phase; (iv) variations among the samples derived from differences in the growth media (such as pheromones, stress, etc.) and (v) nucleosome dynamics across the genome, that will be detected as positional 'fuzziness' in the experimental nucleosomal (48,49).

We present here a systematic analysis of the effect of noise and nucleosome dynamics in defining nucleosome profiles in *Saccharomyces cerevisae*. We have been able to reduce cell-to-cell variability and cell cycle-induced nucleosome dynamics by using carefully synchronized cells and compare them against unsynchronized control cells grown under the same basal conditions. We have then explored the potential sources of variability in nucleosome maps related to different MNase digestion levels (from very mild to very strong) and sequencing procedure (single- versus paired-end). Furthermore, to reduce potential uncertainties related to nucleosome calling algorithms, we have applied a single algorithm with standard defaults, in conjunction with analyses of nucleosome architecture, which are independent of the nucleosome calling algorithm.

Taken together, our findings have allowed us a comprehensive evaluation of nucleosome positioning and stability that may contribute to partially uncover the underlying principles of nucleosome architecture dynamics. The picture derived from our analyses presents navigating nucleosomes along one-dimensional string (i.e. the DNA fiber) that are mainly positioned at specific places in response to strong nucleosome depletion signals generated by either intrinsic properties of DNA, the presence of competing DNA-binding proteins or chromatin-remodeling systems. The proposed model allows a very simple integration of different nucleosome positioning models, and provides clues on the impact of nucleosome fuzziness in the modulation of gene activity.

MATERIALS AND METHODS

Cell-cycle synchronization

Yeast strain BY4741 was grown using fresh YPD media at 30°C until an OD_{600} of 0.2. Then, alpha-factor mating pheromone (GenScript) was added to the culture to a final concentration of 10 μM and the culture was incubated for 2 h to induce cell-cycle arrest in late G1. As a control, an asynchronous culture was grown in parallel to an OD_{600} of 0.8. Both G1-arrest synchronized and asynchronous samples were collected and washed twice with phosphate buffered saline (PBS).

Cell synchrony was monitored by three approaches: flow cytometry (FACS), fluorescence microscopy and budding index calculation. For FACS analysis, cells were fixed with 100% EtOH, spun down and washed once with 1× SSC buffer (150 mM NaCl, 15 mM sodium citrate, pH 7.80). Removal of RNA and proteins were carried out by incubation with RNase A (0, 5 mg/ml, Roche) and proteinase K (0.5 mg/ml, Roche), correspondingly. Samples were briefly sonicated by using the Bioruptor system and mixed with 500 μl SSC buffer containing 0.1 mg/ml propidium iodide (PI, Sigma-Aldrich). Fluorescence emitted from DNA-intercalated PI was measured by Beckman Coulter EPICS® XL flow cytometer.

Cell-cycle phase was also monitored by fluorescence microscopy and budding index calculation. For these purposes, cells were briefly sonicated and fixed with EtOH in a similar manner as for FACS. Fixed cells were then resuspended in 200 μl PBS containing Hoechst stain at 30 μg/ml. Finally, cells were placed on a glass slide and visualized by fluorescence microscopy (Nikon E600 microscope). For budding index calculation, a sample from EtOH-fixed cells was placed on a hemocytometer and visualized under a phase contrast microscope to count the number of budded and unbudded cells.

Nucleosomal DNA extraction

Nucleosomal DNA from G1-arrest and asynchronous samples was prepared as previously described (15). The overall nucleosome digestion was accurately controlled by carrying out several digestion reactions with MNase at concentrations of 0.04, 0.08, 0.12 and 0.16 U, respectively, at 37°C for 30 min. The reactions were stopped by addition of EDTA to a final concentration of 0.02 M and subsequently incubated with RNase A (0.1 mg) for 1 h at 37°C and further treated with Proteinase K at 37°C for 1 h. DNA was extracted using phenol–chloroform extraction and concentrated by ethanol precipitation.

The percentage of mononucleosomal DNA fragments was examined by 2% agarose gels. Furthermore, the integrity and size distribution of digested fragments were determined using the microfluidics-based platform Bioanalyzer (Agilent) prior to DNA sequencing. Typically, samples containing >80% mononucleosomal fragments were sent for sequencing. In addition, over- and under-digested samples were selected based on the proportion of mononucleosomal fragments. Over-digested samples were obtained by 0.12 U MNase digestions, which yielded to only mononucleosomes. Under-digested samples were obtained by 0.04 U of MNase yielding to mono-, di- and tri-nucleosomes.

Nucleosomal DNA sequencing

Libraries of nucleosome fragments were prepared and adapted for deep sequencing using the standard Illumina protocol and sequenced them as single-end paired-end on Genome Analyzer IIx and HiSeq 2000 devices. Data were processed with a standard GA base calling pipeline to convert initial raw images into sequences, as described previously (15). Raw reads are available at the

ENA-SRA website (http://www.ebi.ac.uk/ena) with accession number ERP004019.

RNA Isolation and gene-expression arrays

Cells were collected at the same interval as nucleosomal DNA samples in icy-water and harvested by spinning for 3–4 min at 6000 rpm, frozen in liquid nitrogen and stored at −80°C. Total cellular RNA was extracted using the RNeasy kit (Qiagen), following the manufacturer's instructions with the spheroplasting protocol (0.5 mg/ml zymolase). The total RNA was hybridized to Affymetrix GeneChip Yeast Genome 2.0 arrays for gene-expression analysis. Raw and processed files available in ArrayExpress under accession number E-MTAB-2195.

Data processing and nucleosome calling

Reads from all samples were mapped to yeast genome (SacCer3, UCSC) with Bowtie (50) aligner and imported in R/Bioconductor framework (51). Single-end reads were resized to 50 bp and shifted downstream to align reads mapping in opposite strands using nucleR (52). Paired-end reads were trimmed to 50 bp maintaining the original center. Genome-wide coverage was normalized using the total number of reads in every experiment and scaled by a factor of 10^6 to obtain the units of reads per million (r.p.m.). Peak calling was performed after noise filtering using nucleR parameters: peak width = 125 bp, peak detection threshold = 35%, maximum overlap = 50 bp.

TSS clustering

Using the nucleosome calls obtained previously, we classified every gene according to their nucleosome architecture around the TSS. The closest nucleosome at or immediately downstream TSS was annotated as the +1 nucleosome. The nucleosome immediately upstream of the +1 nucleosome was annotated as the −1 nucleosome. After a visual analysis of the classifications, nucleosome calls were considered as well-positioned (W) when nucleR peak width score (score_w; positioning) and height score (score_h; coverage) were higher than 0.4 and 0.6, respectively. Otherwise, the nucleosome call was considered fuzzy (F). Accordingly with previous observations (39), the NFR was defined as the distance between the dyads of the −1/+1 nucleosome and it was annotated as 'open' if this distance was >215 bp or as 'closed' otherwise. The classification of a given gene was determined by the positioning of the −1 nucleosome, the width of the NFR and the positioning of the +1 nucleosome. Special cases such as when the −1 nucleosome was >300 bp further from the TSS (annotated as M, missing), the −1/+1 nucleosome calls were overlapped or when the regions −300:+300 bp had more than a 25% of uncovered bases were excluded from the analysis.

Evaluation of nucleosome architecture variability between genes

In order to obtain an accurate estimate of gene architecture similarity/dissimilarity between samples, we defined the following metrics. Profile: we considered a gene promoter stable between two samples if Pearson's correlation in the window of −300:300 bp around TSS was >0.7; conversely, we consider variable if the correlation was <0.5. Cluster: we considered a gene promoter stable if cluster dimensions (–1/NFR/+1) matched; otherwise we annotated as a relevant variation of this architecture when two of the clustering dimensions varied between samples. The +1/–1 nucleosome: we considered a stable nucleosome classification if nucleR provided the same classification (with thresholds for score_w and score_h noted above) for the two samples; we considered a relevant change if the nucleosome call was changing in classification and the absolute difference of the aggregated score (nucleR's default score = 0.5*score_h + 0.5*score_w) was >0.25 (which implies a change of at least one quartile of the classification). NFR: we considered it stable if the NFR distance was annotated equally between samples (open/close/overlap/missing); or variable when a change in class was happening and the distance between −1/+1 nucleosomes differed >100 bp. Those genes that do not satisfy any of the criteria are considered out of the stability/variability threshold.

Elastic energy model

Elastic energy was calculated using a mesoscopic model of DNA flexibility (53–56) with parameters derived from molecular dynamics simulations (57) as described previously (15). In short, for every tiled sequence of 147 bp in the genome, we calculated the increment of energy required to pass from a relaxed DNA conformation to a nucleosome-shaped conformation, using an experimental reference structure (58).

Statistical positioning model

The very simple statistical-positioning model featured in this article considers that, after the energetic barrier in the NFR, nucleosomes are arranged statistically with a lineal increasing fuzziness. We decided to simulate a population of nucleosome reads centered in the +1 nucleosome with a dyad deviation of 25 bp. Dyads of downstream nucleosomes (+2, +3, . . .) were spaced 147 + 14 bp (accounting for average linker DNA length) with an increasing deviation of the dyad of +5 bp in every step and a decreasing number of reads equal to the 4% of the previous peak. The dyad of the −1 nucleosome was placed 147+100 bp (247 bp in total) upstream the +1 for the closed NFR and 147+200 bp (347 bp in total) for the open NFR. The following upstream nucleosomes (–2, −3, . . .) were defined as in the case of the downstream model but adjusting the deviation in 35 bp in the −1 nucleosome plus 5 bp in every following step, with a linker length of 18 bp. Different values of the different parameters in the model were selected after a grid search maximizing the correlation of the model with the average experimental distribution.

TFBS prediction

Transcription factor binding sites (TFBSs) were derived from the position weight matrices (PWMs) available in JASPAR database for yeast (59). For every PWM, the

genome-wide binding scores and predicted TFBS were calculated using R/Bioconductor Biostrings library with default parameters. Regions with annotated TFBS were pooled and their coverage was calculated as a measure of global TF affinity genome-wide.

RESULTS

In an attempt to eliminate the noise resulting from cell-population heterogeneity, we have synchronized yeast cultures at the late G1 cell-cycle phase. Two synchronized cultures (Supplementary Figure S1) were considered as biological replicas. We labeled them as replicas 1 and 2 and subsequently isolated their nucleosomal DNA under similar MNase-digestion conditions. The digestion level was determined visually by agarose gel electrophoresis and the microfluidics-based platform Bioanalyzer (Agilent), as shown in Supplementary Figure S2. Accordingly, both samples yielded a major peak ~147 bp corresponding to mononucleosomes, a secondary defined peak ~295 bp corresponding to di-nucleosomes and residual peaks at ~60 bp that might be assigned to either tetrasomes or other DNA–protein complexes.

The effect of single- versus paired-end sequencing

Eventually, replicas 1 and 2 were sequenced both as single-end (1x) and paired-end (2x), assuming a minimal sequencing bias. We obtained high sequencing read depths in all our datasets (fold-coverage ranges between 13X and 177X for individual experiments), which were then processed using nucleR package as described in methods (52). Nucleosome profiles around TSSs were classified based on the positioning of −1 nucleosome (fuzzy (F), well positioned (W) or missing (M)) and +1 nucleosome (F or W) according to their nucleR score (Materials and methods section), and the width of NFR. NFR state was identified as open (typically ~130-bp wide) or closed (~30-bp wide) according to previously reported bimodal distribution (39). We were able to classify ~90% of yeast gene promoters into nucleosome architectures for both replicas (Supplementary Figure S3). The remaining 10% could not be classified due to either low coverage, undefined +1 or overlapping nucleosomes and were not considered in the remaining analysis. We explored the variability of the results based on different assignment criteria, and notably, nucleosome calls appeared quite robust regardless the threshold parameters applied in nucleR (Supplementary Table S1), proving the accuracy of the algorithm and further validating that detected variability does not respond to bioinformatics artifacts.

In principle, since every sample was both single- and paired-end sequenced, nucleosome maps should strongly resemble one another. However, when we compared the nucleosome profiles around TSSs, they displayed different architectures depending on the sequencing method (Figure 1, comparison of A and C with B and D). Single-end sequencing generated noisier nucleosome maps, leading to higher populations of fuzzy nucleosomes. This observation is illustrated in more detail in Supplementary Table S2, which shows the distribution of genes according to classifying parameters such as −1/+1 positioning or NFR width. Conversely, the percentage of −1 or +1 fuzzy nucleosomes is considerably reduced in paired-end sequenced samples, confirming that single-end sequencing yields an artificial enrichment of fuzzy nucleosomes. Interestingly, while the general coverage around TSSs is reasonably preserved (Supplementary Table S3 first column), only ~38–36% of the nucleosome architectures are conserved (Supplementary Table S3, second column) and ~23%of the genes show clearly distinct nucleosome classifications depending on the sequencing (1x versus 2x) method (Supplementary Table S3, seventh column). Intriguingly, when we analyzed the individual parameters describing a nucleosomal architecture, we observed that 63–64% of +1 nucleosome positions have the same annotation in both single- and paired-end sequencing, and only 5–7% display very different annotations. In the case of −1 nucleosome position, the preservation is less pronounced, with 61% conservation (52% for replica 2) and 7% discrepancy (12% for replica 2). Moreover 69% (56% in replica 2) of the NFRs maintain the same annotation, while 6%–9% can show width changes of up to >100 bp.

The effect of biological replica variability

In order to minimize the noise coming from experimental protocols (out of the specific sequencing procedure), we compared the nucleosome maps of the two synchronized replicas generated by paired-end sequencing, which both present a similar coverage of 50X. As shown in Figure 1, nucleosome patterns are quite similar in both replicas, dominated by WoW, WcW and a myriad of families characterized by a +1W and −1F/M (Figure 1, Supplementary Figure S3). However, when we analyzed individual genes (Supplementary Table S3), significant differences arise between replicas. Nearly 90% of genes show reasonably similar coverage profiles, but only 48% of them maintain their nucleosome architecture in the two replicas. Even though the majority of changes are subtle, usually only affecting one nucleosome position (e.g. W→F or F→M), ~15% of genes show significant changes in class annotation. Interestingly, only 3% of genes show dramatic changes in NFR in contrast to 6–7% that show different −1 and +1 nucleosome localizations (Supplementary Table S3), indicating that NFR is more conserved and less prone to variations than flanking nucleosome positions.

Taken together, our findings show that inter-replica variations are not negligible and point out that nucleosome positioning is intrinsically highly plastic and dynamic. In accordance with this observation, elastic energy models propose that a 10-bp sliding of a nucleosome would face a general energy barrier (i.e. the difference between the best and worse wrapping configuration) of ~13 kcal/mol in average (<1 kcal/mol for sliding one single position), while in a larger scale, we would find mesoscopic barriers of ~47 kcal/mol (Supplementary Figure S4), which might contribute to phasing. Similarly, atomistic molecular dynamics simulations detect spontaneous sliding of one base step at the

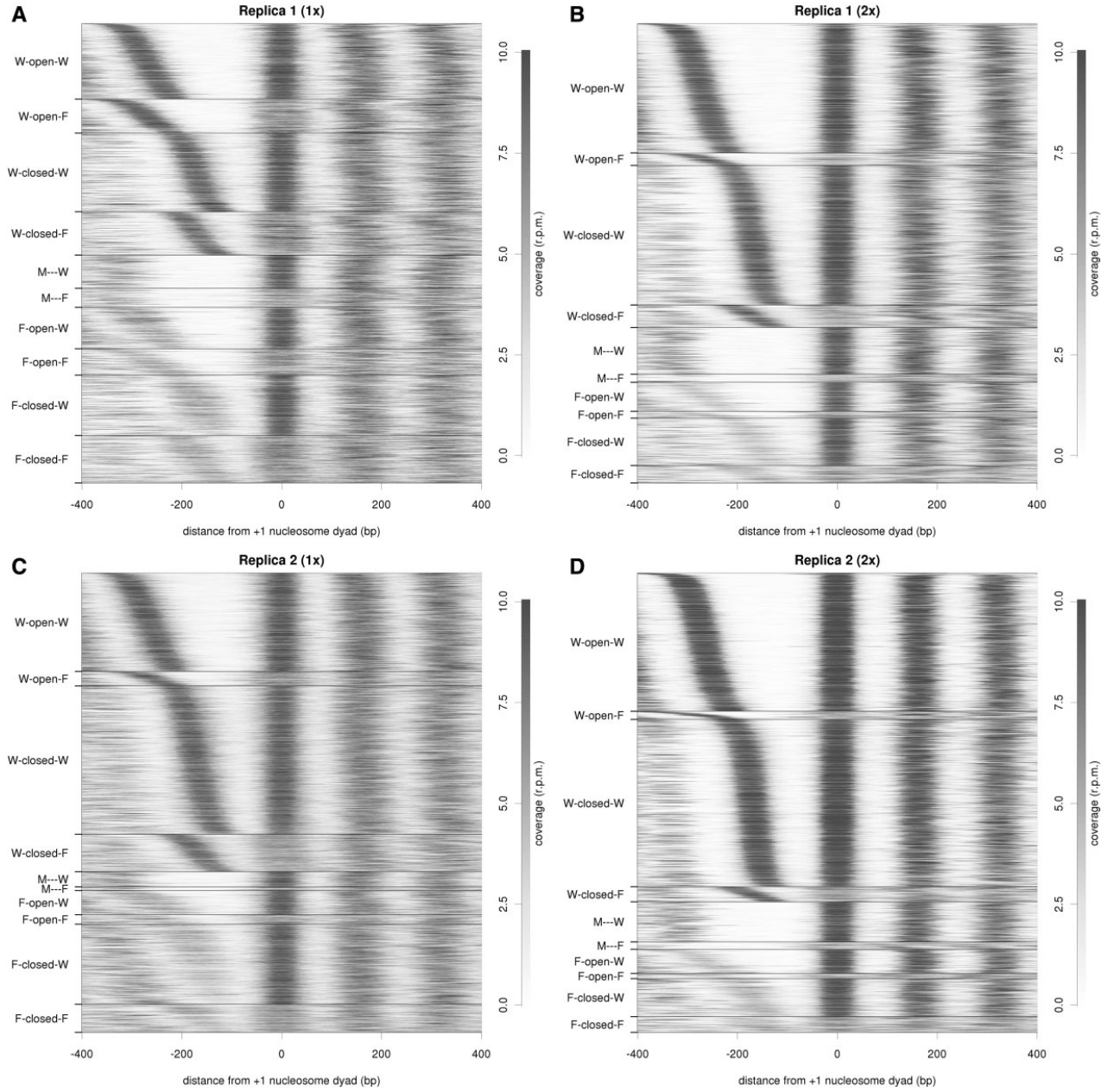


Figure 1. Nucleosome coverage and gene clustering in single- and paired-end sequencing. Heat maps showing nucleosome occupancy around TSS in replicas 1 (top) and 2 (bottom) for single-end sequencing (1x, left) and paired-end sequencing (2x, right). Genes are clustered based on their nucleosome profile and their coverage is plotted taking +1 nucleosome dyad as ′0′.

multi-nanosecond time scale (60). Consequently, with such a dynamic organization, nucleosomes tend to change positions constantly and hence, well-positioned nucleosomes might not actually be intrinsically tightly placed in the absence of negative 'nucleosome depletion' signals. Indeed, the deformation energy required to wrap fuzzy or well-positioned nucleosomes is similar (~200 kcal/mol for a 145-bp double-stranded DNA, Supplementary Figure S5). Fuzziness is then likely to be the default state for nucleosomes in a random DNA in the absence of additional factors, such as NFR.

The effect of cell diversity

In order to determine the variability caused by cell heterogeneity, we included an asynchronous yeast culture in our analysis, labeled as 'asynchronous' sample (Supplementary Figure S1). This sample also produced

well-defined nucleosomal maps, where 94% of yeast genes had well-located +1 nucleosomes and around 87% of promoters could be assigned to a particular nucleosomal architecture around TSS.

The asynchronous sample shows clear differences when compared to the synchronized replicas 1 and 2, as shown in Figures 1 and 2 and Supplementary Table S1. WoW and WcW nucleosome classes decrease while −1 F/M or +1 F nucleosome positions are more prevalent in asynchronous maps (proportion test P-value $< 2.2 \times 10^{-16}$), suggesting that cell-cycle progression may induce chromatin rearrangements around TSSs that may be reflected as diffuse nucleosome signals in MNase-Seq experiments derived from asynchronous samples.

Interestingly, analysis of individual genes provides a clearer impact of cell heterogeneity-dependent variability. Differences in nucleosome coverage profiles and nucleosome architectures are higher when the asynchronous sample is compared against biological replicas than when biological replicas are directly compared (Supplementary Table S2, columns 6 and 7). As anticipated from Figure 2, the increase in −1 and +1 nucleosome fuzziness seems to be the main responsible for variations in chromatin structure around TSSs (Supplementary Tables S1 and S2). In average 475 genes pass from WoW or WcW in rep1/2 to some fuzzy structure in the asynchronous replica. As a result the ratio well-positioned/non-well-positioned −1 nucleosomes decreases from 1.8–2.3 (replicas 1 and 2) to 1.4 (in asynchronous) (proportion test for replica 1 P-value $= 1.07 \times 10^{-10}$, for replica 2 P-value $< 2.2 \times 10^{-16}$), and similarly in case of +1 nucleosomes, going from 5.0–7.0 (replicas 1 and 2) to 3.7 (proportion test for replica 1 P-value $= 5 \times 10^{-11}$, for replica 2 P-value $< 2.2 \times 10^{-16}$). Conversely, NFR width between synchronized and asynchronous samples shows similar changes as between biological replicas, suggesting that cell cycle-dependent chromatin rearrangements do not lead to massive nucleosome eviction around TSSs, which would dramatically alter NFR dimensions (proportion test for replica 1 P-value $= 0.22$, for replica 2 P-value $= 0.09$).

Overall, our detailed comparison of cell cycle synchronized and asynchronous samples reveals that asynchronous experiments contain an additional source of noise due to the cell cycle-dependent nucleosome dynamics, and that caution is necessary with maps derived from asynchronous samples (those typically available in the literature), since average maps can mask the existence of two populations with completely different nucleosomal architectures. Fuzzy nucleosome signals can be in reality the result of a mixed population where some cells contain well-localized nucleosomes while others are nucleosome-depleted at the same region, or rather that a well-localized nucleosome has moved to a neighboring position in different cells (Supplementary Figure S6). Indeed, we observed that 105 genes displayed very similar coverage profiles around TSSs between biological replicas (Pearson's correlation > 0.7), but clearly differed with the asynchronous sample (Pearson's correlation < 0.5). Of note, although GO and pathway annotation analyses were not able to find any particular

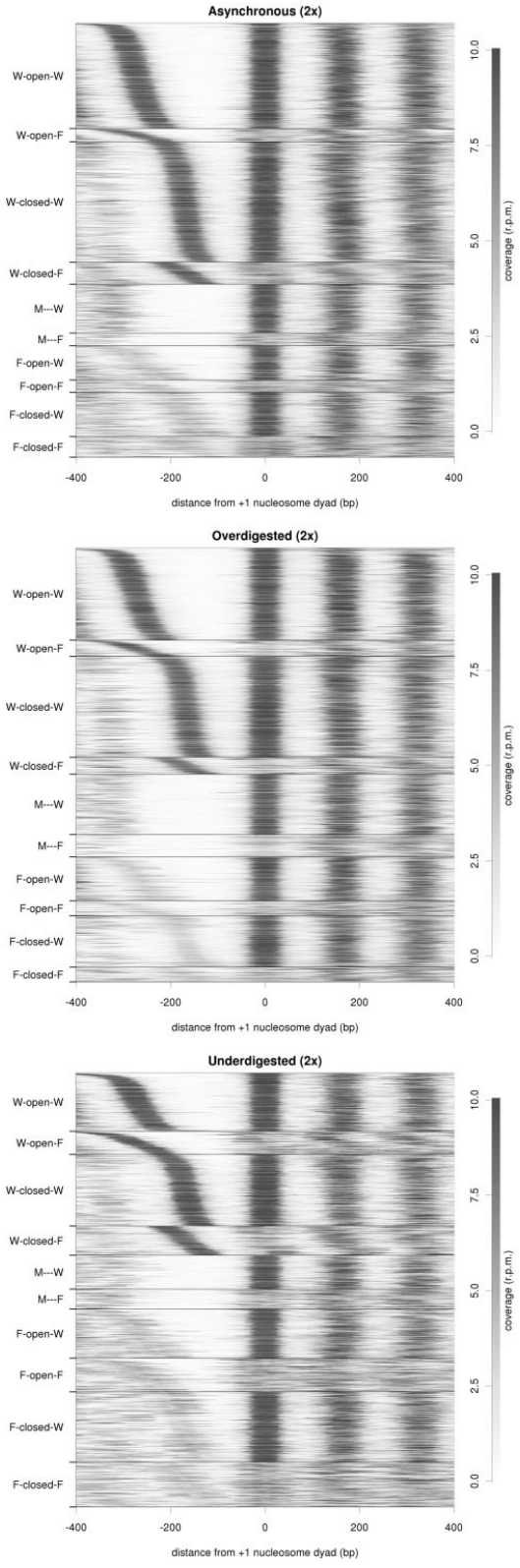


Figure 2. Nucleosome coverage and clustering under different experimental conditions. Similar to Figure 2, but for asynchronous (top), over-digested (middle) and under-digested (bottom) samples.

enrichment in this set of genes, 15 of them are annotated as cell cycle periodic genes in Cyclebase (61), which represents a small enrichment in cell-cycle related functions (from 9.8% in genomic mean to 14%; proportion test P-value = 0.087). Interestingly, cell cycle-related genes belong to the G1/S regulon rather than to the alpha-factor pathway, indicating that the source of variation does not derive from alpha-factor stimuli. A more detailed analysis would require data for every individual cell stages, but nonetheless, the present results support the notion that chromatin reorganization might be coupled to cell-cycle regulation mechanisms.

The effect of MNase digestion

To investigate the bias effect of MNase digestion on the generation of nucleosome maps (a quite ignored source of variability), we have used two additional MNase-Seq experiments derived from a G1-synchronized culture but treated under either more aggressive (over-digested sample) or milder (under-digested sample) MNase-digestion conditions. The samples were sequenced using paired-end technology and a similar data processing for a direct comparison with other replicas.

As shown in the Bioanalyzer histograms (Supplementary Figure S2), over-digestion of chromatin leads to the disappearance of the di-nucleosome signal and to a broader mono-nucleosome peak that is shifted towards shorter fragments, probably caused by certain intra-nucleosomal cleavage. Nucleosome architectures of the over-digested sample are well-defined, with a clear +1 nucleosome signal in 95% of the gene promoters and unambiguously assigned nucleosome families, similar to replicas 1 and 2. Yet, the analysis of nucleosome pattern distributions reveals clear differences between the over-digested sample and replicas 1 and 2 (Figure 2, Supplementary Table S2). In over-digested chromatin, the prevalence of canonical nucleosome classes (i.e. WoW and WcW) decreases and fuzzy −1 nucleosomes are enriched. Thus, ∼800 genes move from WoW/WcW patterns in replicas 1 and 2 to a fuzzier configuration in the over-digested sample (proportion test P-value $< 2.2 \times 10^{-16}$). The number of missing −1 nucleosomes increases from 711–510 (replicas 1 and 2) to 1154 (proportion test P-value $< 2.2 \times 10^{-16}$). Overall, these findings suggest that excessive MNase digestion can lead to partial degradation of some well-positioned nucleosomes, resulting in fuzzier nucleosome peaks, or even to the complete disassociation of unstable nucleosomes, leading to loss of certain nucleosome signals. This behavior is further confirmed with a higher mean deviation of the nucleosome dyad position in the over-digested sample (Kolmogorov–Smirnov test P-value $= 2 \cdot 10^{-6}$) (Supplementary Figure S7).

We have further explored the impact of over-digestion on nucleosomal architectures by the analysis of individual genes (Supplementary Table S3). While only 15% of the genes show clearly different nucleosome arrangements between replicas, up to 25% (18% in replica 2) change with respect to the over-digested sample. Similarly, the variability in −1 and +1 nucleosome annotations also increases from 6–7% to 10–13%. In contrast, NFR width seems to be very resistant to digestion conditions, pointing out that a more aggressive digestion mostly result in partial degradation of nucleosomes but rarely in their complete eviction around TSSs (Supplementary Tables S2–S3 and Figure 2).

On the other hand, the under-digested sample exhibits well-defined mono-, di-, tri- and even tetra-nucleosomal peaks (Supplementary Figure S2). Intriguingly, paired-end sequencing only yielded 78% of gene coverage and ∼69% of TSSs could be classified into nucleosome families, whereas the classification was 86.5 ± 3.7% in previous experiments. The significantly (proportion test P-value $= 8 \times 10^{-9}$ for replica 1, P-value $= 2 \times 10^{-16}$ for replica 2) enrichment of depleted areas might account for the under-representation of longer fragments in the sequencing reactions, since it is well established that deep sequencing favors the amplification of short over long fragments (62). Moreover, we observed a higher frequency of overlapping nucleosomes (10.2% in over-digested in comparison with 4.6% and 2.5% in replicas 1 and 2, proportion test P-value $< 2.2 \times 10^{-16}$), which might correspond in reality to DNA in complex with other proteins. Interestingly, the uncovered regions specific to under-digested sample are distributed over the entire genome without any significant enrichment according to GO analysis. Yet, they show a clear preference for AT-rich segments (3.38 higher fold) and intergenic regions (15% enrichment over background, simulated P-value $< 10^{-5}$).

Despite the poor ability of sequencing protocols to deal with long fragments, we were able to recover several long reads analogous to di-nucleosomal signals, being ∼3% of them longer than 300 bp in the under-digested sample. In contrast, we only rescued 0.4% of these long reads in the over-digested sample. Therefore, the presence of di-nucleosome signals introduce another source of noise in the nucleosome maps, since nucleosome calling algorithms usually consider peak signals as mono-nucleosomes and align them on the genome based on their middle position, which is assumed to correspond to the nucleosome dyad. However, the mid-point of di-nucleosomes is actually located on the linker region, leading to a counter-phase location with respect to mono-nucleosome signals. This observation is shown in Figure 3, where short-, mid- or long-sized fragments of over- and under-digested samples are compared. The counter-phasing is more explicit in under-digestion, which contains more di-nucleosome derived signals and thus, leads to a higher noise in terms of linker length and nucleosome phasing (Figure 3B).

Overall, our observations show that MNase digestion levels may strongly bias nucleosome maps by intra-nucleosomal cleavage or the introduction of longer internucleosomal or non-nucleosomal protected regions. Caution is then necessary, since MNase is not a mere spectator of nucleosome architecture, but bias it in one way or the other, introducing a noise that need to be considered when maps obtained under different degradation conditions are compared. This warning is especially important since MNase is an enzyme whose activity is not always easy to control.

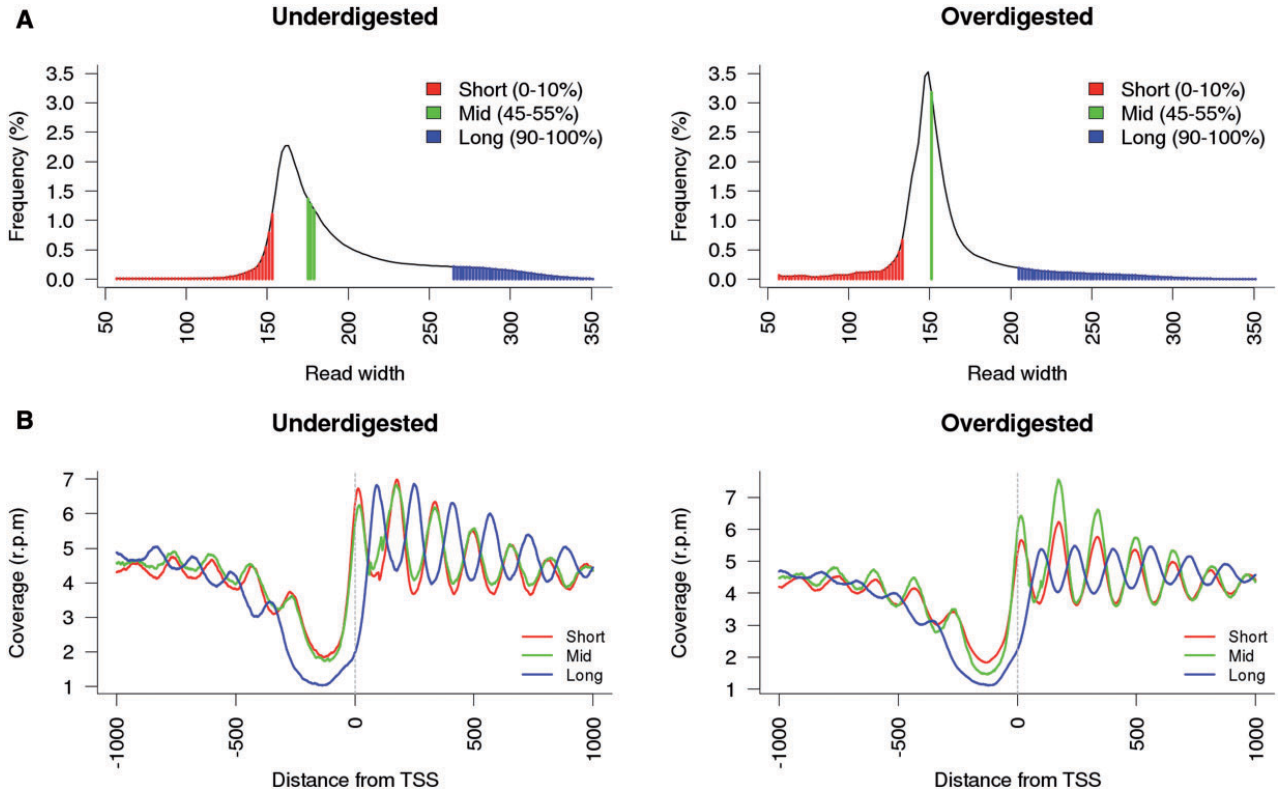


Figure 3. Effect of variable read length on map coverage. (**A**) Coverage distribution of short, mid and long reads in under-digested (left) and over-digested (right) samples. (**B**) Normalized coverage profiles of trimmed reads around TSSs derived from different sequencing lengths.

The effect of sequencing read depth

Interestingly, when comparing paired-end experiments with extremely large read depth (for example over-digested (2x) and asyncronized (2x) datasets, with a coverage of 146X and 177X, respectively), the nucleosomal maps do not show better agreement than when comparing datasets with lower read depth (such as single-end replicas 1 and 2, with coverage ~13X and 45X, respectively). Certainly, low read-depth can increase the noise in nucleosome maps but it does not seem to be a dramatic contributor in the present article.

The effect of basal expression level

In order to assess the possible effect of noise due to different transcription rates, we measured the absolute levels of gene expression in our G1 synchronized samples (replicas 1 and 2) using expression arrays (see Materials and methods section). Based on the hybridization results, we then selected the top 500 and the bottom 500 genes according to their normalized expression level and analyzed in detail the nucleosome organization around TSSs. Lowly expressed genes display better defined nucleosome organization than highly expressed genes (with ~8% more robust coverage profiles; Wilcoxon rank sum test P-value < 0.0002). Notably, this differential organization is increased to 13% when gene body is considered (Wilcoxon rank sum test P-value $< 2.2e-16$).

Underlying factors in nucleosome positioning

Heterogeneity in MNase-Seq experiments may be the responsible of the low accuracy in the available nucleosome positioning predictive models when other than the training datasets are employed, challenging the validation of nucleosome positioning models. In contrast, our systematic analysis under controlled conditions allowed us to obtain a robust set of nucleosome profiles (showing a correlation >0.7 and displaying the same nucleosome architecture in biological replicas). This set (comprising 3096 genes) represents the well-conserved nucleosome architectures in late G1 cell-cycle phase and reveals that WcW (1306 genes) and WoW (1164 genes) are the most pervasive classes, followed by M–W (263 genes) and FcF (155 genes) classes. We can use then these maps to evaluate the goodness of different predictive models, without being exposed to noise and uncertainties in the experimental data.

We compared the experimental WcW and WoW nucleosome maps with an extremely simple statistical model (see Materials and methods section) that places nucleosomes every 161–165 bp (147 + 14 – 18 bp due to linker variation), starting from NFR with decreasing nucleosome phasing (Figure 4A). Once the NFR is defined, the model is able to explain the majority of nucleosome positioning around TSSs, further confirming that nucleosome location in TSSs is largely determined by a barrier, i.e.

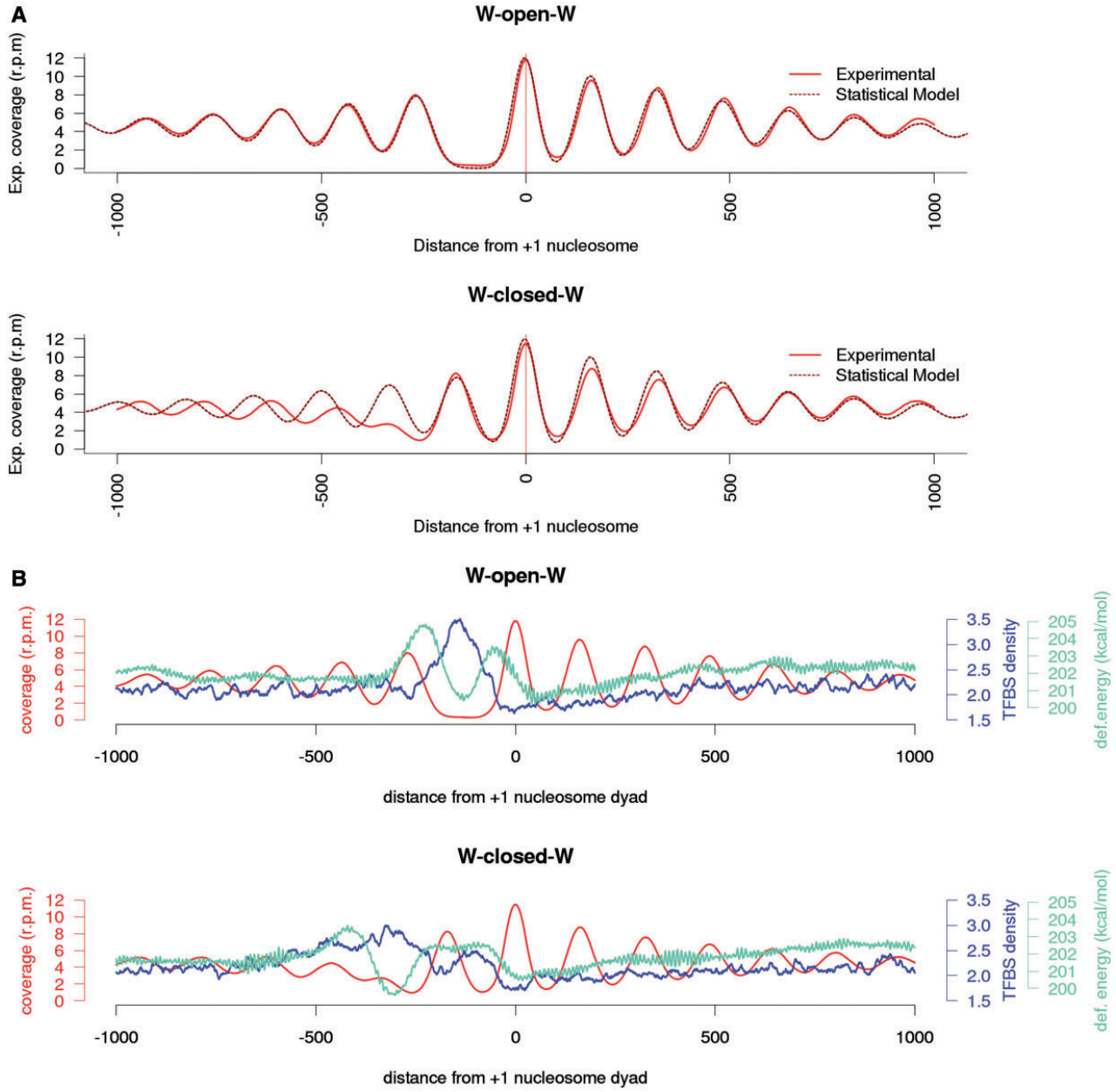


Figure 4. Statistical positioning and intrinsic DNA energetic barriers. (**A**) Average experimental nucleosome coverage from WoW (top) and WcW (bottom) patterns in replica 1 are compared against a nucleosome positioning statistical model. (**B**) The experimental coverage of WoW (top) and WcW (bottom) classes (red) are overlapped with deformation energy (cyan) and predictive TFBS (blue) around TSSs.

NFR, which subsequently places arrays of nucleosomes with decay in positioning as they separate from the NFR signal. It seems then that the crucial step to position nucleosomes is to determine the placement of the NFR. To analyze the determinants of NFR position we studied TFBS and compute the deformation energy required to wrap DNA around the histone core (see 'Materials and Methods' section). As shown in Figure 4B, NFRs at WoW architectures present intrinsically different DNA properties leading to an anomalously large deformation energies at NFR. This suggests that physical properties can define the boundaries off the NFRs in this family, but they also erroneously predict a well-positioned nucleosome in the middle of the NFR. It is clear (Figure 4B) that competition with TFs avoids the binding of the nucleosome in the middle of the NFR. Clearly, TF binding is then crucial in determining the integrity of NFRs and hence the phasing of the nucleosome arrays in WoW architectures. The synergetic effect of physical properties and TFBS is also clear in nucleosome placements in the WcW family, where the region around TSS is marked by an unusual profile of physical properties

and a distinct pattern of TFBSs. The strong +1 nucleosome signal fits perfectly in a region of low cost for wrapping DNA around a nucleosome and depleted in TFBS.

Taken together, our analysis of robust nucleosome maps suggest that simple statistical positioning is a key responsible of the basal nucleosome positioning, with NFR being the main signal organizing nucleosome string. Physical properties and binding to effector proteins, such as TFs, act in a synergistic manner to define NFR location and boundaries and to define then the phasing of the nucleosome string. Clearly, the chromatin remodelers will act on this basal activity to facilitate positioning of nucleosomes in a regularly spaced array.

Lastly, we analyzed the heterogeneity in nucleosome architecture along coding regions, taking into account that they do not show the typical TSS nucleosome pattern (see Supplementary Tables S4–S5). Interestingly, our findings indicate that the conclusions obtained for TSS regions are also valid for gene body regions. Clearly, the heterogeneity and plasticity in nucleosome architecture is not a differential property of the TSS vicinities, since also coding regions show a significant degree of flexibility that surely reflects the intrinsic mobility of nucleosomes, especially for highly expressed genes.

DISCUSSION

Many genome-wide nucleosome maps are now available in the literature for model organisms, but the correlation among the aligned maps is typically poor. This variability is a first indication that nucleosomes are not as rigidly placed as suggested from single dataset analyses. However, it is difficult to determine at what extent such variability is due to biological sources (for instance action of chromatin remodelers) or related to experimental procedures (such as MNase different activity rate).

In an attempt to remove, as much as possible, experimental and sequencing biases, we have addressed several sources of noise that might lead to artifactual nucleosome fuzziness in MNase-Seq derived nucleosome maps (Figure 5). Single- and paired-end sequencing of identical samples demonstrates that single-end sequencing generates a high level of fuzziness, prompting to low reproducible nucleosome maps and challenging accurate analyses of nucleosome arrangements along the DNA fiber. Furthermore, when identical samples are treated under different MNase digestion conditions, high MNase levels lead to intra-nucleosomal cleavage and partial degradation of some well-positioned, but not necessarily very stable, nucleosomes. Since the two ends of the nucleosome might not be equally protected due to differences in DNA–histone interactions (63,64), over-digestion can lead to unsymmetrical degradation of nucleosome, which could be reflected in nucleosome maps as an increase in the overall fuzziness. On the other hand, milder digestion can also lead to severe noise, since the poor sequencing of long reads might cause loss of information about those regions enriched in long fragments. Moreover, the alignment of di-nucleosomal signals can yield to counter-phased positioning with respect to mono-nucleosomal signals, generating noise in the nucleosome maps, especially in the linker regions.

Computational processing of raw data is another source of 'spurious noise'. Converting read-alignment information into nucleosome positions implies the assumption of some predictive models (especially for single-end sequencing reads). Furthermore, the classification of nucleosome signals also requires the use of arbitrary thresholds. We have minimized these potential factors by applying various classification thresholds in the nucleR nucleosome calling algorithm and by the use of orthogonal metrics such as direct read information. Our findings suggest that the detected variability in our analyses do not respond to computational artifacts. Additionally, the observed variability is not a consequence of poor sequencing coverage either, since no reduction of variability is found when samples with relatively high and low coverage are compared.

Up to date, most of the available MNase-Seq derived nucleosome maps originate from asynchronous populations. This asynchrony can produce considerable amount of noise due to cell cycle-dependent nucleosome changes which may be related to gene-expression alterations or chromatin compaction and relaxation. Therefore these maps actually reproduce average nucleosome phasing in different populations, some of which might be stable and well-positioned in one cell while more diffuse in others. Such variations can generate misleading nucleosome maps that might not accurately reproduce the actual chromatin organization, and in fact only reflect the statistical averaging of different cell populations. Consistently, our comparative analysis between synchronized and asynchronous samples shows that asynchrony contributes to an increase of ~30% in nucleosome fuzziness. Regions with striking chromatin alterations along cell cycle are usually detected as regions with fuzzy and not well-defined nucleosomes in the usual asynchronous maps. Interestingly, highly active genes in a certain cell-cycle stage show in general less-conserved nucleosome distributions, connecting gene activity and nucleosome mobility.

We have tried to further minimize the biological noise level by generating nucleosome maps from cell cycle synchronized biological replicas, in order to have a reliable detection and accurate analysis of nucleosome positioning. Although average nucleosome maps are identical, differences at the gene level are not negligible. Thus, even if general nucleosome profiles are reasonably conserved, only 30% of the well-positioned nucleosomes are located at the same genomic place (within ±5-bp deviation) in two biologically equivalent replicas. Our results suggest that in principle nucleosomes are mobile along the DNA fiber, since intrinsic sliding barriers are small. This can lead to a kind of 'spread' of signal along average position, or alternatively, to the population of different nucleosomal arrangements in different cells, giving in all cases to a fuzzy signal in nucleosomal maps.

The localization pattern that we (and many others before) have found obeys at great extent to a simple statistical positioning, where nucleosome arrays are placed

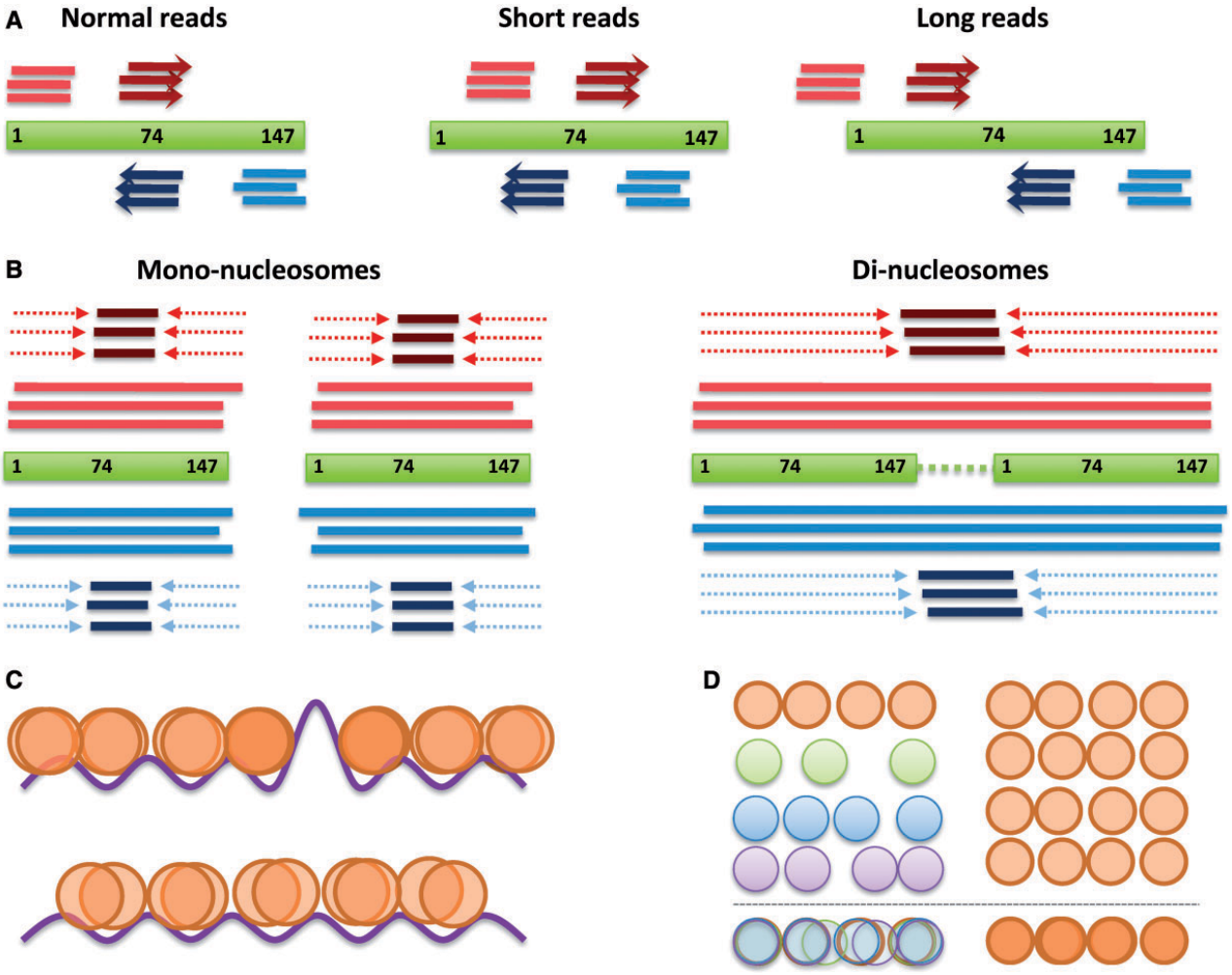


Figure 5. Possible sources of intrinsic nucleosome noise. (**A**) In single-end sequencing, reads mapped in opposite strands (light red, light blue) are shifted 74-bp downstream to align the nucleosome dyad (dark red, dark blue). Despite this approach is suitable for mono-nucleosome fragment alignment (left), shorter (middle) or longer fragments (right) are misaligned, causing a fuzzy peak coverage. (**B**) In paired-end sequencing, the detection of mono-nucleosome dyads can be obtained by trimming the reads (left). However, in the case of long di-nucleosome fragments, the trimmed reads are aligned to the linker space between mono-nucleosomes, which in turn increase the fuzziness. (**C**) Energetic barriers due to intrinsic DNA deformability potential or presence of competing proteins (represented as purple line) act as a phasing element in adjacent nucleosomes (top) leading to well-localized nucleosome signals. In the absence of such barriers, the periodicity of this potential cannot act in nucleosome phasing (bottom) leading to diffuse signals originated by spontaneous nucleosome sliding. (**D**) Individual nucleosome arrays of asynchronous cells in different stages of the cell-cycle capture intrinsic chromatin dynamics (left) which is visualized as fuzzy signals. This effect is minimized in synchronized cell populations (right).

starting from NFRs (which are defined with a small noise level and a high reliability in biological replicas), with a lineal decrease in positioning as separated from the NFR. NFR borders are marked by unusual DNA physical properties that hampers wrapping of DNA around histone core (*in vitro* and *in vivo*) and by regions which are highly prevalent in TFBSs (*in vivo*). Under normal physiological conditions, sequence-encoded physical properties and protein binding (including TFBSs) act synergistically to define the NFR and accordingly, the nucleosome array. Such synergy can be enhanced or reduced by chromatin remodelers, which are shown to facilitate nucleosome positioning by packing them against a barrier, like NFRs (18,31). This could partially explain the differences often found between *in vivo* and *in vitro* nucleosomal maps.

Taken together, our robust set of nucleosome profiles have enabled us to carefully inspect the various sources of noise and dissociate them from the actual nucleosome dynamics in the cell, which in all cases are otherwise captured as positional 'fuzziness' in the nucleosome maps. Finally, our systematic approach provides an insight into nucleosome positioning determinants and may guide the standardization of MNase-Seq experiments in order to generate reproducible genome-wide nucleosome patterns. Finally, our results shed light in the requirement of discarding the current static picture of nucleosome positioning and start considering nucleosomes

as intrinsically mobile entities navigating along the DNA fiber.

ACCESSION NUMBERS

Raw reads are available at the ENA-SRA website (http://www.ebi.ac.uk/ena) with accession number ERP004019. Raw and processed expression data for G1 stage is available in ArrayExpress platform under accession number E-MTAB-2195.

SUPPLEMENTARY DATA

Supplementary Data are available in at NAR Online.

FUNDING

Spanish Ministry of Science and Innovation [BIO2012-32868]; Instituto de Salud Carlos III (INB); European Research Council (SimDNA project). Funding for open access charge: Spanish Ministry of Science and Innovation [BIO2009-10964 and Consolider E-Science], Instituto de Salud Carlos III (INB-Genoma España and COMBIOMED RETICS) and Fundación Marcelino Botín.

Conflict of interest statement. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

1. Kornberg,R.D. and Lorch,Y. (1999) Twenty-five years of the nucleosome, review fundamental particle of the eukaryote chromosome. *Cell*, **98**, 285–294.
2. Malik,H.S. and Henikoff,S. (2003) Phylogenomics of the nucleosome. *Nat. Struct. Biol.*, **10**, 882–891.
3. 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 A resolution. *Nature*, **389**, 251–260.
4. Tolstorukov,M.Y., Volfovsky,N., Stephens,R.M. and Park,P.J. (2011) Impact of chromatin structure on sequence variability in the human genome. *Nat. Struct. Mol. Biol.*, **18**, 510–515.
5. Tilgner,H., Nikolaou,C., Althammer,S., Sammeth,M., Beato,M., Valcárcel,J. and Guigó,R. (2009) Nucleosome positioning as a determinant of exon recognition. *Nat. Struct. Mol. Biol.*, **16**, 996–1001.
6. Workman,J.L. and Buchman,A.R. (1993) Multiple functions of nucleosomes and regulatory factors in transcription. *Trends Biochem. Sci.*, **18**, 90–95.
7. Jiang,C. and Pugh,B.F. (2009) A compiled and systematic reference map of nucleosome positions across the Saccharomyces cerevisiae genome. *Genome Biol.*, **10**, R109.
8. 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.
9. Lee,C.-K., Shibata,Y., Rao,B., Strahl,B.D. and Lieb,J.D. (2004) Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nat. Genet.*, **36**, 900–905.
10. Lee,W., Tillo,D., Bray,N., Morse,R.H., Davis,R.W., Hughes,T.R. and Nislow,C. (2007) A high-resolution atlas of nucleosome occupancy in yeast. *Nat. Genet.*, **39**, 1235–1244.
11. Sadeh,R. and Allis,C.D. (2011) Genome-wide ''re''-modeling of nucleosome positions. *Cell*, **147**, 263–266.
12. Segal,E., Fondufe-Mittendorf,Y., Chen,L., Thåström,A., Field,Y., Moore,I.K., Wang,J.-P.Z. and Widom,J. (2006) A genomic code for nucleosome positioning. *Nature*, **442**, 772–778.
13. Zhang,L., Ma,H. and Pugh,B.F. (2011) Stable and dynamic nucleosome states during a meiotic developmental process. *Genome Res.*, **21**, 875–884.
14. Albert,I., Mavrich,T.N., Tomsho,L.P., Qi,J., Zanton,S.J., Schuster,S.C. and Pugh,B.F. (2007) Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. *Nature*, **446**, 572–576.
15. Deniz,Ö., Flores,O., Battistini,F., Perez,A., Soler-Lopez,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.
16. Rach,E.A., Winter,D.R., Benjamin,A.M., Corcoran,D.L., Ni,T., Zhu,J. and Ohler,U. (2011) Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level. *PLoS Genet.*, **7**, e1001274.
17. Radman-Livaja,M., Verzijlbergen,K.F., Weiner,A., van Welsem,T., Friedman,N., Rando,O.J. and van Leeuwen,F. (2011) Patterns and mechanisms of ancestral histone protein inheritance in budding yeast. *PLoS Biol.*, **9**, e1001075.
18. 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*, **332**, 977–980.
19. Wang,X., Bai,L., Bryant,G.O. and Ptashne,M. (2011) Nucleosomes and the accessibility problem. *Trends Genet.*, **27**, 487–492.
20. Kornberg,R. (1981) The location of nucleosomes in chromatin: specific or statistical? *Nature*, **292**, 579–580.
21. 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.
22. Field,Y., Kaplan,N., Fondufe-Mittendorf,Y., Moore,I.K., Sharon,E., Lubling,Y., Widom,J. and Segal,E. (2008) Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS Comput. Biol.*, **4**, e1000216.
23. Miele,V., Vaillant,C., D'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.
24. Morozov,A.V., Fortney,K., Gaykalova,D.A., Studitsky,V.M., Widom,J. and Siggia,E.D. (2009) Using DNA mechanics to predict in vitro nucleosome positions and formation energies. *Nucleic Acids Res.*, **37**, 4707–4722.
25. Lantermann,A.B., Straub,T., Strålfors,A., Yuan,G.-C., Ekwall,K. and Korber,P. (2010) Schizosaccharomyces pombe genome-wide nucleosome mapping reveals positioning mechanisms distinct from those of *Saccharomyces cerevisiae*. *Nat. Struct. Mol. Biol.*, **17**, 251–257.
26. Ioshikhes,I., Hosid,S. and Pugh,F. (2011) Variety of genomic DNA patterns for nucleosome positioning. *Genome Res.*, **21**, 1863–1871.
27. Scipioni,A. and De Santis,P. (2011) Predicting Nucleosome Positioning in Genomes: Physical and Bioinformatic Approaches. *Biophys. Chem.*, **155**, 53–64.
28. Trifonov,E.N. (2011) Cracking the chromatin code: Precise rule of nucleosome positioning. *Phys. Life Rev.*, **8**, 39–50.
29. 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.
30. 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.
31. Yen,K., Vinayachandran,V., Batta,K., Koerber,R.T. and Pugh,B.F. (2012) Genome-wide nucleosome specificity and directionality of chromatin remodelers. *Cell*, **149**, 1461–1473.
32. Struhl,K. and Segal,E. (2013) Determinants of nucleosome positioning. *Nat. Struct. Mol. Biol.*, **20**, 267–273.
33. Tsankov,A., Yanagisawa,Y., Rhind,N., Regev,A. and Rando,O.J. (2011) Evolutionary divergence of intrinsic and trans-regulated

nucleosome positioning sequences reveals plastic rules for chromatin organization. *Genome Res.*, **21**, 1851–1862.
34. Wang,X., Bryant,G.O., Floer,M., Spagna,D. and Ptashne,M. (2011) An effect of DNA sequence on nucleosome occupancy and removal. *Nat. Struct. Mol. Biol.*, **18**, 507–509.
35. 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.
36. Segal,E. and Widom,J. (2009) What controls nucleosome positions? *Trends Genet.*, **25**, 335–343.
37. Hihara,S., Pack,C.-G., Kaizu,K., Tani,T., Hanafusa,T., Nozaki,T., Takemoto,S., Yoshimi,T., Yokota,H., Imamoto,N. *et al.* (2012) Local nucleosome dynamics facilitate chromatin accessibility in living mammalian cells. *Cell Rep.*, **2**, 1645–1656.
38. Möbius,W., Osberg,B., Tsankov,A.M., Rando,O.J. and Gerland,U. (2013) Toward a unified physical model of nucleosome patterns flanking transcription start sites. *Proc. Natl Acad. Sci. USA*, **110**, 5719–5724.
39. Zaugg,J.B. and Luscombe,N.M. (2011) A genomic model of condition-specific nucleosome behaviour explains transcriptional activity in yeast. *Genome Res.*, **22**, 84–94.
40. Vavouri,T. and Lehner,B. (2011) Chromatin organization in sperm may be the major functional consequence of base composition variation in the human genome. *PLoS Genet.*, Apr; **7**, e1002036.
41. Arya,G., Maitra,A. and Grigoryev,S.A. (2010) A structural perspective on the where, how, why, and what of nucleosome positioning. *J. Biomol. Struct. Dyn.*, **27**, 803–820.
42. 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**, 887–898.
43. Huebert,D.J., Kuan,P.-F., Keleş,S. and Gasch,A.P. (2012) Dynamic changes in nucleosome occupancy are not predictive of gene expression dynamics but are linked to transcription and chromatin regulators. *Mol. Cell. Biol.*, **32**, 1645–1653.
44. Bai,L. and Morozov,A.V. (2010) Gene regulation by nucleosome positioning. *Trends Genet.*, **26**, 476–483.
45. Kuan,P.F., Huebert,D., Gasch,A. and Keles,S. (2009) A non-homogeneous hidden-state model on first order differences for automatic detection of nucleosome positions. *Stat. Appl. Genet. Mol. Biol.*, **8**:Article 29.
46. Tsankov,A.M., Thompson,D.A., Socha,A., Regev,A. and Rando,O.J. (2010) The role of nucleosome positioning in the evolution of gene regulation. *PLoS Biol.*, **8**, e1000414.
47. Ozonov,E.A. and van Nimwegen,E. (2013) Nucleosome free regions in yeast promoters result from competitive binding of transcription factors that interact with chromatin modifiers. *PLoS Comput. Biol.*, **9**, e1003181.
48. Belch,Y., Yang,J., Liu,Y., Malkaram,S.A., Liu,R., Riethoven,J.J. and Ladunga,I. (2010) Weakly positioned nucleosomes enhance the transcriptional competency of chromatin. *PLoS One 24*, **5**, e12984.
49. Lehner,B. (2010) Conflict between noise and plasticity in yeast. *PLoS Genet.*, **6**, e1001185.
50. 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.
51. Gentleman,R.C., Carey,V.J., Bates,D.M., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.
52. Flores,O. and Orozco,M. (2011) nucleR: a package for non-parametric nucleosome positioning. *Bioinformatics*, **27**, 2149–2150.
53. Olson,W.K. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl Acad. Sci.*, **95**, 11163–11168.
54. Lankas,F., Sponer,J., Langowski,J. and Cheatham,T.E. (2003) DNA basepair step deformability inferred from molecular dynamics simulations. *Biophys. J.*, **85**, 2872–2883.
55. Lavery,R., Zakrzewska,K., Beveridge,D., Bishop,T.C., Case,D.A., Cheatham,T., Dixit,S., Jayaram,B., Lankas,F., Laughton,C. *et al.* (2010) 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**, 299–313.
56. 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**, 2379–2394.
57. Pérez,A., Marchán,I., Svozil,D., Sponer,J., Cheatham,T.E., Laughton,C.A. and Orozco,M. (2007) Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J.*, **92**, 3817–3829.
58. Davey,C.A., Sargent,D.F., Luger,K., Maeder,A.W. and Richmond,T.J. (2002) Solvent Mediated Interactions in the Structure of the Nucleosome Core Particle at 1.9Å Resolution†. *J. Mol. Biol.*, **319**, 1097–1113.
59. Bryne,J.C., Valen,E., Tang,M.-H.E., Marstrand,T., Winther,O., da Piedade,I., Krogh,A., Lenhard,B. and Sandelin,A. (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.*, **36**, D102–D106.
60. Portella,G., Battistini,F. and Orozco,M. (2013) Understanding the connection between epigenetic DNA methylation and nucleosome positioning from computer simulations. *PLoS Comput. Biol.*, Nov; **9**, e1003354.
61. Gauthier,N.P., Jensen,L.J., Wernersson,R., Brunak,S. and Jensen,T.S. (2010) Cyclebase.org: version 2.0, an updated comprehensive, multi-species repository of cell cycle experiments and derived analysis results. *Nucleic Acids Res.*, **38**, D699–D702.
62. Dabney,J. and Meyer,M. (2012) Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *Biotechniques*, **52**, 87–94.
63. Luger,K., Rechsteiner,T.J., Flaus,A.J., Waye,M.M. and Richmond,T.J. (1997) Characterization of nucleosome core particles containing histone proteins made in bacteria. *J. Mol. Biol.*, **272**, 301–311.
64. Luger,K. (2006) Dynamic nucleosomes. *Chromosome Res.*, **14**, 5–16.

2.3. Hog1 bypasses stress-mediated down-regulation of transcription by RNA polymerase II redistribution and chromatin remodeling

This article was the outcome of the collaboration with the group of Prof. Francesc Posas of the University Pompeu Fabra (Barcelona, Spain). Here, we studied a peculiar effect in chromatin organization caused by osmostress in Hog1-regulated genes.

Despite the specific topic covered here is far from our expertise, our recent publication of nucleR package (page 109) enabled us to help in the processing and analysis of the MNase-seq experiments performed in this work

Hog1 is a stress-activated protein kinase which acts as a key regulator of the transcription in response to osmolarity changes in yeast. In general, drastic environmental changes induce a down-regulation of the gene expression, together with an up-regulation of the stress-response genes. In this article, we studied how Hog1-related genes under stress show a co-localization of this protein with RNA Polymerase II, linked to an increment in the expression of those genes in comparison with constitutively expressed genes. Additionally, those changes in the expression reveal a drastic chromatin remodeling in Hog1-related genes, despite chromatin structure is not altered in a global scale (Fig. 26). Furthermore, the effect of Hog1 in these changes was validated with a Hog1-mutant strain of yeast which. When Hog1 was knock-out, Hog1-related genes described previously did not show the effects above, pointing that the presence of Hog1 is crucial for an adequate response to stress.

Publication:

Nadal-Ribelles, M. *, Conde, N. *, Flores, O., González-Vallinas, J., Eyras, E., Orozco, M., de Nadal, E., Posas, F. (2012) "*Hog1 bypasses stress-mediated down-regulation of transcription by RNA polymerase II redistribution and chromatin remodeling*". Genome Biology 13: R106.

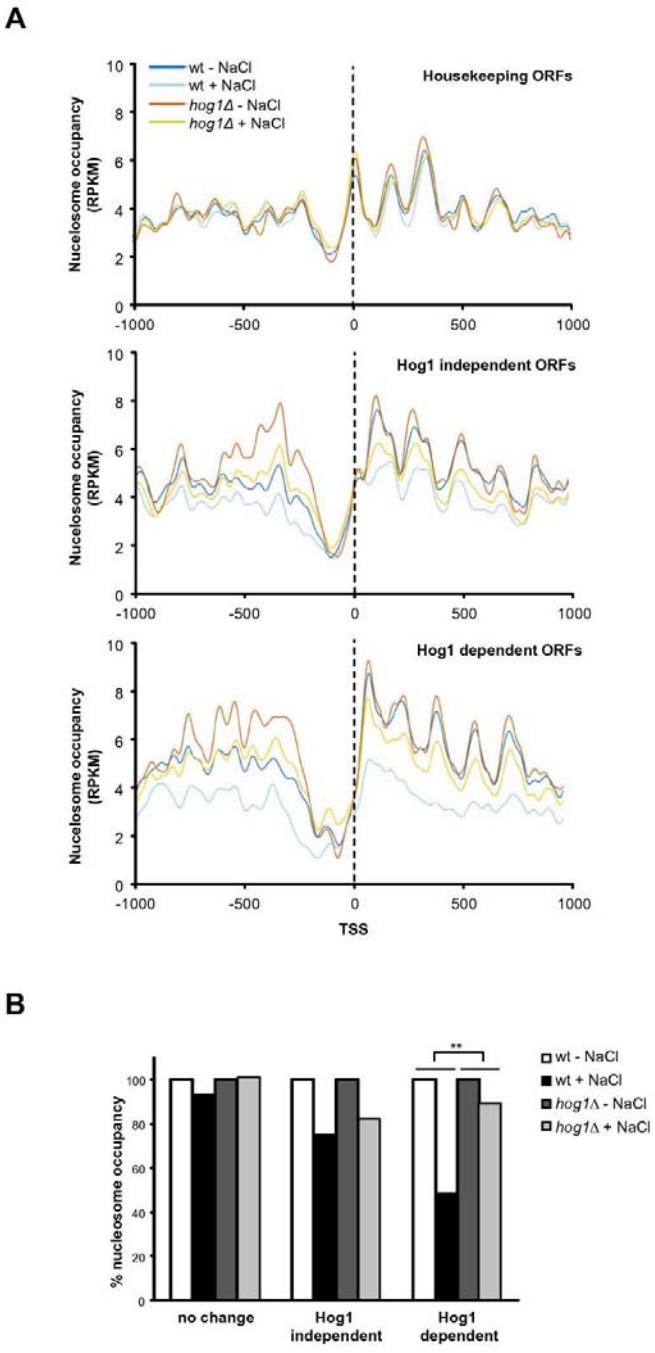


Fig. 26. Effects of the chromatin alteration due to osmostress and hog1 regulation. A) Nucleosome coverage around TSS of housekeeping (top), hog1 independent (middle) and hog1 dependent (bottom) genes for wild-type and hog1Δ mutant with (+ NaCl) and without osmostress (- NaCl). B) The changes in the chromatin attributed to hog 1-dependent gens under osmostress are significantly higher than in the rest of the cases.

Nadal-Ribelles et al. Genome Biology 2012, 13:R106
http://genomebiology.com/2012/13/11/R106

Genome Biology

RESEARCH Open Access

Hog1 bypasses stress-mediated down-regulation of transcription by RNA polymerase II redistribution and chromatin remodeling

Mariona Nadal-Ribelles[1†], Núria Conde[1†], Oscar Flores[2], Juan González-Vallinas[3], Eduardo Eyras[3,4], Modesto Orozco[2], Eulàlia de Nadal[1*] and Francesc Posas[1*]

Abstract

Background: Cells are subjected to dramatic changes of gene expression upon environmental changes. Stress causes a general down-regulation of gene expression together with the induction of a set of stress-responsive genes. The p38-related stress-activated protein kinase Hog1 is an important regulator of transcription upon osmostress in yeast.

Results: Genome-wide localization studies of RNA polymerase II (RNA Pol II) and Hog1 showed that stress induced major changes in RNA Pol II localization, with a shift toward stress-responsive genes relative to housekeeping genes. RNA Pol II relocalization required Hog1, which was also localized to stress-responsive loci. In addition to RNA Pol II-bound genes, Hog1 also localized to RNA polymerase III-bound genes, pointing to a wider role for Hog1 in transcriptional control than initially expected. Interestingly, an increasing association of Hog1 with stress-responsive genes was strongly correlated with chromatin remodeling and increased gene expression. Remarkably, MNase-Seq analysis showed that although chromatin structure was not significantly altered at a genome-wide level in response to stress, there was pronounced chromatin remodeling for those genes that displayed Hog1 association.

Conclusion: Hog1 serves to bypass the general down-regulation of gene expression that occurs in response to osmostress, and does so both by targeting RNA Pol II machinery and by inducing chromatin remodeling at stress-responsive loci.

Background

Yeast cells undergo major changes of gene expression in response to stress [1]. Global gene expression in response to osmostress in yeast has been studied in detail [2-9]. Major changes of gene expression occur in response to stress; many genes are down-regulated together with the up-regulation of a set of stress-responsive genes.

Activation of the high osmolarity glycerol (HOG) pathway upon stress regulates many aspects of cell physiology, including gene expression. The p38-related Hog1 stress-activated protein kinase (SAPK) is the master protein for reprogramming gene expression in response to osmostress through different specific transcription factors [5,7]. Hog1 is recruited to the osmoresponsive genes by these specific factors [10-17]. Once bound to chromatin, Hog1 serves as a platform to recruit RNA polymerase II (RNA Pol II) [13] and associated transcription factors [12,18-20]. Hog1 is present also at the coding regions of stress-responsive genes [15-17], where the kinase is essential for increased association of RNA Pol II and efficient mRNA production in response to osmostress [17]. Moreover, nucleosome positioning of specific stress-responsive loci is altered dramatically in a Hog1-dependent manner through the chromatin structure remodeling (RSC) complex upon osmostress [21].

Here, we assessed the genome-wide enrichment of RNA Pol II and Hog1 in response to stress by chromatin immunoprecipitation (ChIP) followed by sequencing (ChIP-Seq) as well as the re-organization of nucleosomes at stress-responsive loci by micrococcal nuclease followed

* Correspondence: eulalia.nadal@upf.edu; francesc.posas@upf.edu
† Contributed equally
[1]Cell Signaling Unit, Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra (UPF), E-08003 Barcelona, Spain
Full list of author information is available at the end of the article

by sequencing (MNase-Seq). We define a comprehensive picture of the genome-wide regulatory organization of the genome in response to stress and reveal Hog1 as the key protein needed to coordinate RNA Pol II relocalization, chromatin re-organization and osmospecific gene expression.

Results and discussion

Stress induces a rapid recruitment of RNA Pol II at stress-responsive loci

Analyses of gene expression have shown there is a rapid and strong induction of a set of stress-responsive genes in response to stress [2-9]. We quantified the increased fold induction of gene expression of 662 stress-responsive genes from microarray analysis (Materials and methods) and found an overall 6.4-fold increase of gene expression upon osmostress (0.4 M NaCl for 10 minutes). The induction pattern of these osmoresponsive genes in other stress conditions, such as heat shock (15 minutes at 37°C), oxidative stress (320 mM H_2O_2, 30 minutes), protein folding (250 mM dithiothreitol, 60 minutes) and amino acid starvation (30 minutes) [2] showed that osmoresponsive genes display a different expression pattern depending on each stress. In general, there is a poor overlap among the different stresses, with heat and osmostress overlapping the most (32%; Figure S1 in Additional file 1).

Whilst osmostress-induced genes showed a clear induction upon stress, the overall transcription of the whole genome, excluding the set of osmostress-induced genes, showed a 0.16-fold reduction in gene expression upon stress (Figure 1a). These data are consistent with previous reports [9] and indicate there must be a specialized mechanism that permits specific gene expression when global down-regulation of gene expression occurs.

To characterize how the changes of gene expression in response to osmostress are accomplished, we analyzed genome-wide binding of RNA Pol II in response to osmostress by ChIP-Seq in wild-type and *hog1* cells. Association of RNA Pol II with ORFs is reduced when the overall genome is considered, whereas it clearly increases for stress-responsive genes (Figure 1b). Earlier studies showed that some housekeeping genes suffered a strong reduction of RNA Pol II occupancy at early time points in response to stress [22]. This is exemplified by the increase of RNA Pol II at the *STL1* osmoresponsive gene in contrast to the reduction of overall RNA Pol II observed at the housekeeping gene *PMA1*, which encodes an essential H-ATPase (Figure 1c). Thus, *STL1* and *PMA1* genes are clear examples that represent the trend in osmoresponsive versus housekeeping genes. It also shows that while RNA Pol II is lost at the housekeeping genes in both wild-type and *hog1* strains, the wild-type strain shows a faster recovery of RNA Pol II. Of note, the down-regulation of RNA Pol II in house-keeping genes precedes the recruitment of RNA Pol II at stress-responsive genes, indicating that the overall reduction of RNA Pol II occupancy cannot be due to a decrease in its availability. Taken together, genome-wide RNA Pol II localization suggests a strong bias for its localization towards stress-responsive genes.

Stress-responsive genes can be classified into two groups, Hog1-dependent and Hog1-independent, on the basis of gene expression data (Materials and methods; Additional file 2). When the 100 most responsive genes of each group were analyzed we found that indeed there was a clear difference in the degree of induction; the Hog1-dependent genes displayed a fold change that was almost four times higher compared to Hog1-independent genes (Figure S2 in Additional file 1). We then analyzed each group with regard to RNA Pol II association and found that RNA Pol II was recruited significantly to Hog1-dependent (green dots) and Hog1-independent (red dots) responsive genes in response to stress (Figure 1d, left-hand panels). By contrast, recruitment of RNA Pol II to Hog1-dependent genes was significantly different between wild-type and *hog1* strains (green dots), while no differences were observed between a wild-type and a *hog1* strain with regard to Hog1-independent genes (red dots) (Figure 1d, right-hand panels). The association of RNA Pol II with down-regulated genes or housekeeping genes in a wild-type strain was similar to that in a *hog1* strain, indicating that Hog1 does not play a role in the initial changes observed in non-stress-dependent genes (Figure S3 in Additional file 1). Thus, Hog1 has a crucial role in the genome-wide redistribution of RNA Pol II to stress-responsive genes upon stress.

Hog1 associates with the chromatin of RNA Pol II and Pol III genes

Genome-wide studies using ChIP and microarray analysis have been instrumental in uncovering the presence of Hog1 associated with a number of stress-responsive genes as well as its localization at both promoter and coding regions of stress-responsive genes. However, the number of genes uncovered by these approaches has been rather limited and never totaled more than 70 genes [15-17]. The relevance of Hog1 in gene expression and RNA Pol II recruitment suggested that the number of genes with Hog1 association could have been underestimated. We undertook ChIP-Seq analysis to improve the sensitivity of detection and found that Hog1 is present in at least 340 genome loci upon osmostress (0.4 M NaCl for 5 minutes; Figure S4a in Additional file 1). We analyzed binding at 5 minutes because this was the peak of Hog1 association with *STL1* and *CTT1* [19] (Figure S5 in Additional file 1). Albeit ChIP experiments generate data that are population averages, previous single cell analyses showed that Hog1 is activated in all cells similarly upon osmostress and that transcriptional induction correlates very well with the

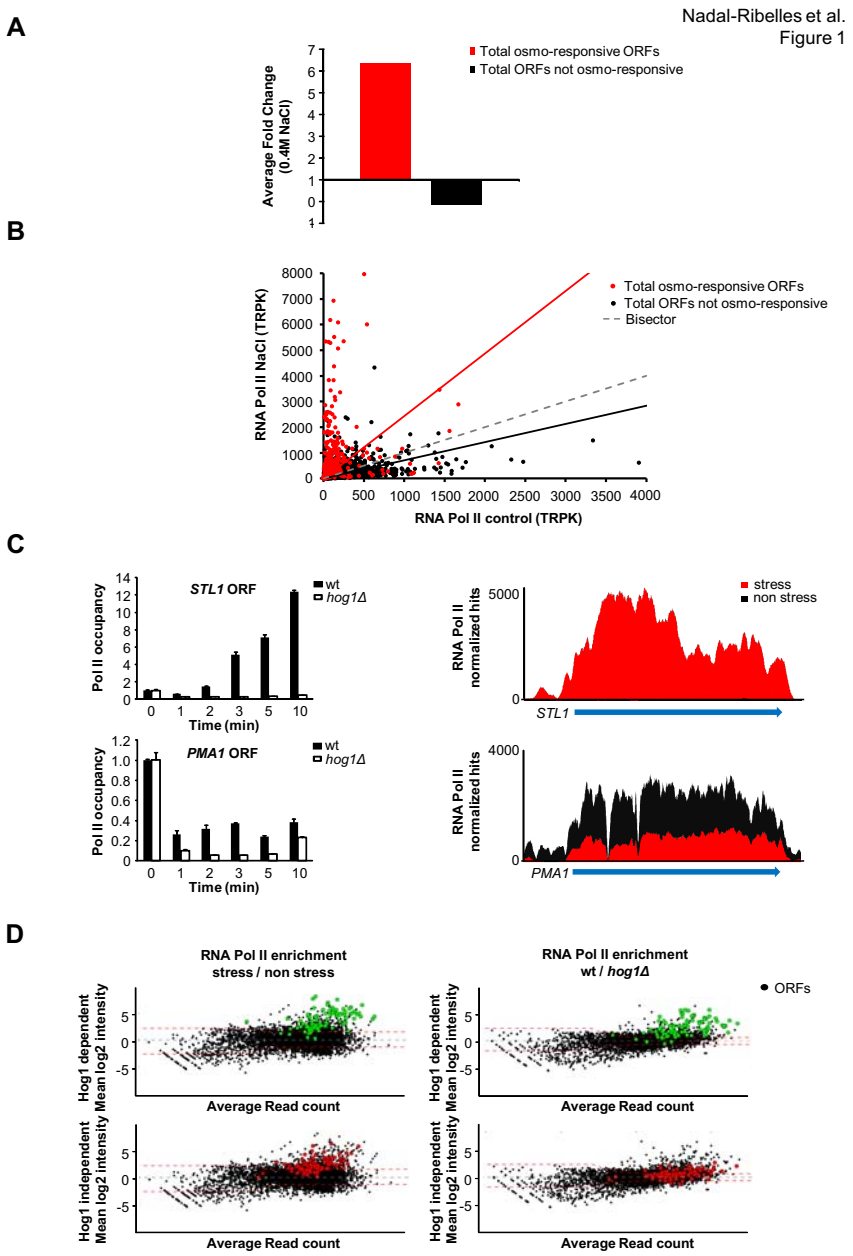


Figure 1 Selective redistribution of RNA Pol II in response to osmostress. **(a)** Comparison of gene expression changes from microarray data on wild-type (wt) cells subjected to osmostress (0.4 M NaCl, 15 minutes; see Materials and methods; Additional file 2). Bars represent the mean fold change for osmoresponsive ORFs with fold change > 1.75 upon osmostress (662 genes) and total ORFs present in the array except those considered osmoresponsive (5,655 genes; see Materials and methods). **(b)** A scatter plot showing RNA Pol II occupancy in basal conditions (YPD; x-axis) versus osmostress (0.4 M NaCl, 10 minutes; y-axis). Each dot represents normalized hits (TRPK; trimmed mean of *M* values normalized read/kilobase density). Black dots and the trend line represent TRPK distribution of total ORFs in the genome except osmoresponsive genes. Osmoresponsive genes are represented as red dots. **(c)** RNA Pol II binding kinetics to osmoresponsive and constitutively expressed genes. Left-hand panels: association of RNA Pol II upon osmostress (0.4 M NaCl, for the indicated times) was assessed by ChIP to *STL1* osmoresponsive gene and *PMA1*. Real-time quantitative PCR (qPCR) results are shown as fold induction of treated versus non-treated (time zero). Right-hand panels: overlapping ChIP-Seq tracks representing RNA Pol II normalized hits at the *STL1* and *PMA1* loci in the presence (red histogram) or in the absence (black histogram) of osmostress. Red and black histograms have been overlaid. The blue arrow indicates annotated ORF. **(d)** Role of Hog1 in RNA Pol II recruitment. *MA* plots of RNA Pol II binding in wild-type upon osmostress (left-hand panel; see Materials and methods). *MA* plots of RNA Pol II binding wild-type versus *hog1* mutant stressed as before. The dotted red line delimits the threshold for significance ($P = 0.0001$). Highlighted dots indicate a subset of 100 osmoresponsive genes that are differentially expressed based on the dependency of the SAPK (Hog1-dependent in green and Hog1-independent in red).

localization of Hog1 to stress-responsive genes [23]. Recruitment of Hog1 was not restricted to RNA Pol II transcribed genes but was present, albeit to a lesser extent, at RNA Pol III transcribed genes as well as long terminal repeat (LTR) DNA regions.

When the presence of Hog1 was analyzed on RNA Pol II transcribed genes, we found that Hog1 was associated with approximately 80% of genes, with expression described to be highly dependent on the SAPK, confirming that Hog1 is widely associated with Hog1-regulated genes. By contrast, only 30% of the genes induced upon osmostress are Hog1-independent and showed Hog1 associated with their loci (Figure 2a; Figure S4b in Additional file 1, green and red dots). Hog1 was not present at down-regulated or house-keeping genes (Figure S4b in Additional file 1, blue and yellow dots); therefore, Hog1 is associated with stress-responsive RNA Pol II genes.

Hog1 has been shown to associate with promoters and the ORFs of stress-responsive genes. We asked whether there was a biased interaction towards promoters or ORFs and found that Hog1 was associated with both promoters and ORFs in a significant number of genes (more than 41%). Also, for those genes where only one region was above the threshold, association of Hog1 with ORFs was more prominent than with promoters (Figure S6 in Additional file 1). When we analyzed association of Hog1 with Hog1-dependent or Hog1-independent genes, we found that Hog1 binding is biased slightly towards the ORFs in Hog1-dependent genes, whereas for genes for which Hog1 was less relevant, Hog1 localization was associated mainly with promoters (Figure 2b). Several scenarios could explain the presence of the SAPK at Hog1-independent loci, such as the use of a too stringent threshold for Hog1 dependency, their presence close to a Hog1-dependent gene, or the induction of the gene is mediated by redundant pathways, including Hog1. Thus, stress-induced RNA Pol II transcribed genes appear to have strong enrichment of Hog1 at their promoters and ORFs.

Remarkably, Hog1 was also associated with RNA Pol III transcribed genes, including at least 16 tRNA genes as well as the two reference genes *SCR1* and *RPR1* (Figure 2c). ChIP experiments showed similar kinetics of association of Hog1 with tF(GAA)D, *RPR1* and *SCR1* as with RNA Pol II transcribed genes. We then investigated the association of RNA Pol III (Rpc82 subunit) with two tRNA loci (tF(GAA)D and tP(UGG)O3) and found that, albeit RNA Pol III dissociated rapidly from chromatin in a stress-dependent manner, a rapid recovery of RNA Pol III levels occurred in wild-type that was not observed in *hog1* cells (Figure S7 in Additional file 1).

It has been reported that Hog1 interacts with RNA Pol II (most likely through Rpb1), which facilitates gene expression in RNA Pol II transcribed genes [13,17]. Therefore, we used co-precipitation experiments in extracts from cells expressing endogenously tagged HA-Hog1 and Myc-Rpc82 (a subunit of the RNA Pol III complex not shared with RNA Pol II) to assess whether Hog1 is able to interact with RNA Pol III. We found Hog1 was able to interact with endogenous tagged-Rpc82 and vice versa (Figure 2d). It is worth noting that this interaction was observed only when cells were subjected to osmostress. Thus, Hog1 is targeted to RNA Pol III loci in response to stress and associates physically with RNA Pol III, as it does with RNA Pol II transcribed genes.

Efficient recruitment of RNA Pol II and maximal gene expression requires Hog1

To assess the relevance of the association of Hog1 with RNA Pol II, we compared the degree of gene expression of stress-responsive genes with the presence of RNA Pol II and Hog1 (Figure 3a). Several groups of osmoresponsive genes can be identified, depending on the presence of Hog1 and/or RNA Pol II upon stress. A group of genes showed no significant association of Hog1 with RNA Pol II but were induced upon osmostress. This group of genes correlated quite well with genes that showed stabilization of mRNAs upon stress (40 out of the 43 analyzed were found to be stabilized) [8,9]. We found a significant number of genes with increased RNA Pol II association that did not have Hog1 present on them (91 out of 391); these genes correspond to Hog1-independent genes. There was, however, a prominent overlapping group of genes that showed increased expression, increased recruitment of RNA Pol II and association with Hog1 (a total of 144 genes; Figure 3a). We then compared the degree of gene induction in those groups of genes and found that there was a strong correlation between the presence of both Hog1 and RNA Pol II with high expression rates when compared to genes that were enriched only with RNA Pol II (Figure 3b). If the presence of Hog1 improved the recruitment of RNA Pol II and transcription, it should be possible to establish a quantitative relationship between weak, or strong, Hog1 binding with RNA Pol II and transcription. Co-localization studies of Hog1 with RNA Pol II showed that RNA Pol II association with stress-responsive genes was more efficient for genes with higher Hog1 association (Figure 3c). Therefore, a high level of induction in stress-responsive genes is accomplished by strong association with Hog1 and increased RNA Pol II recruitment.

It has been reported that very low salt (0.1 M NaCl) stress results in maximal Hog1 activation. When cells are exposed to higher concentrations of NaCl, however, activation of Hog1 remains associated with stress-responsive loci for an extended period of time [19,23]. If the presence of Hog1 improves the recruitment of RNA Pol II and transcription efficiency, it would be expected that an increase of Hog1 at specific promoters will result in enhanced

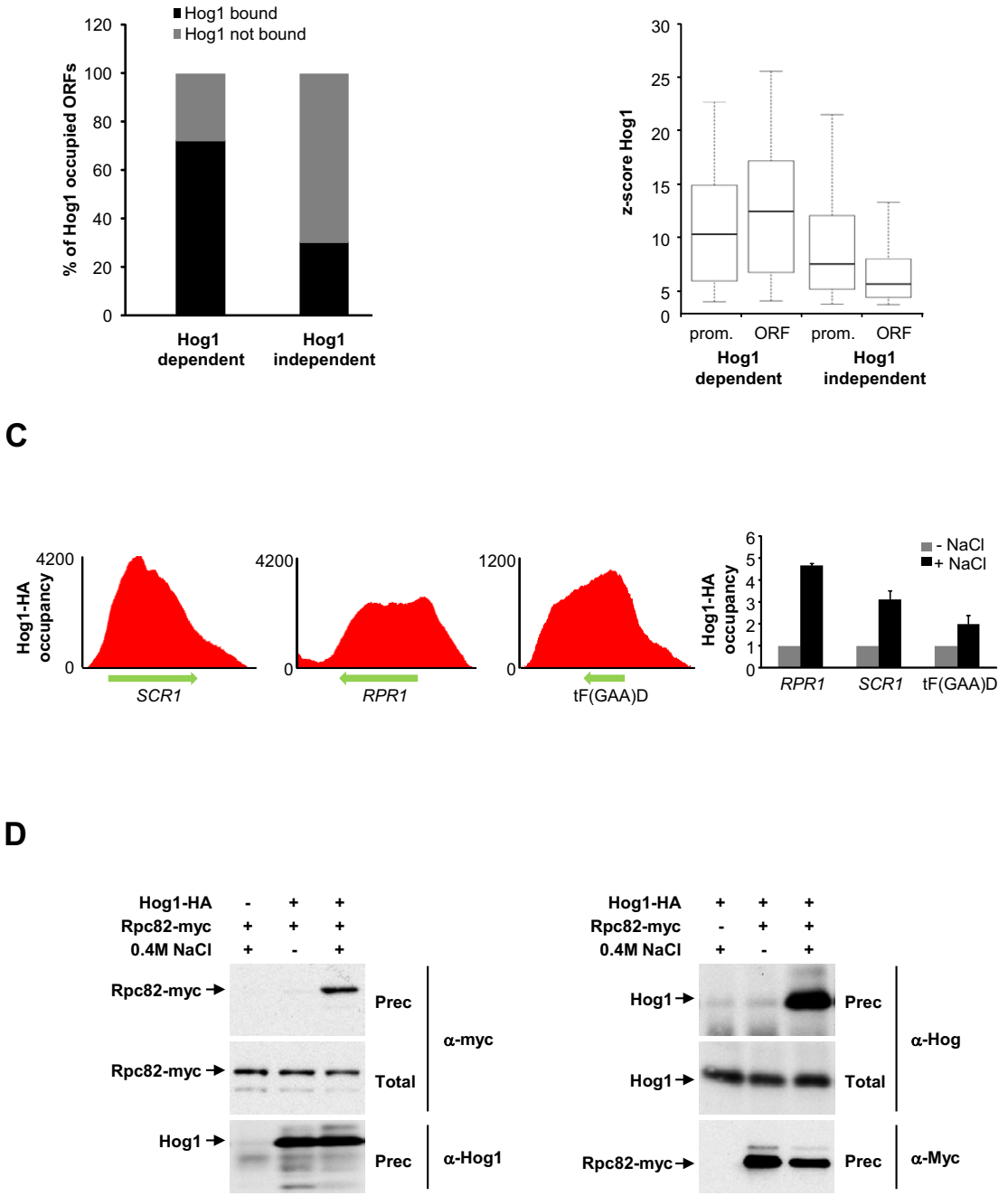


Figure 2 Hog1 associates with RNA Pol II and RNA Pol III loci upon stress. **(a)** Stacked bars representing the percentage of Hog1-occupied ORFs upon stress as determined by ChIP-Seq data for Hog1-dependent and -independent subsets of genes (*Z*-score 4; $P = 0.0001$), as in Figure 1d). **(b)** Boxplot of Hog1 association upon stress (*Z*-score) at promoters (from -500 to the transcription start site) and ORFs in genes whose expression is classified as Hog1-dependent or Hog1-independent. **(c)** Hog1 recruitment at RNA Pol III loci of *SCR1*, *RPR1* and tF(GAA)D. Histograms represent Hog1 ChIP-Seq profiles from a non-stressed sample normalized to that obtained upon osmostress. Right-hand panel: Validation of Hog1 recruitment to novel Hog1 sites by ChIP. **(d)** *In vivo* binding of Hog1 and RNA Pol III. Hog1-HA and a specific subunit of RNA Pol III (Rpc82-Myc) were expressed from their endogenous locus. Samples were taken before (-) and after (+) stress (0.4 M NaCl, 10 minutes).

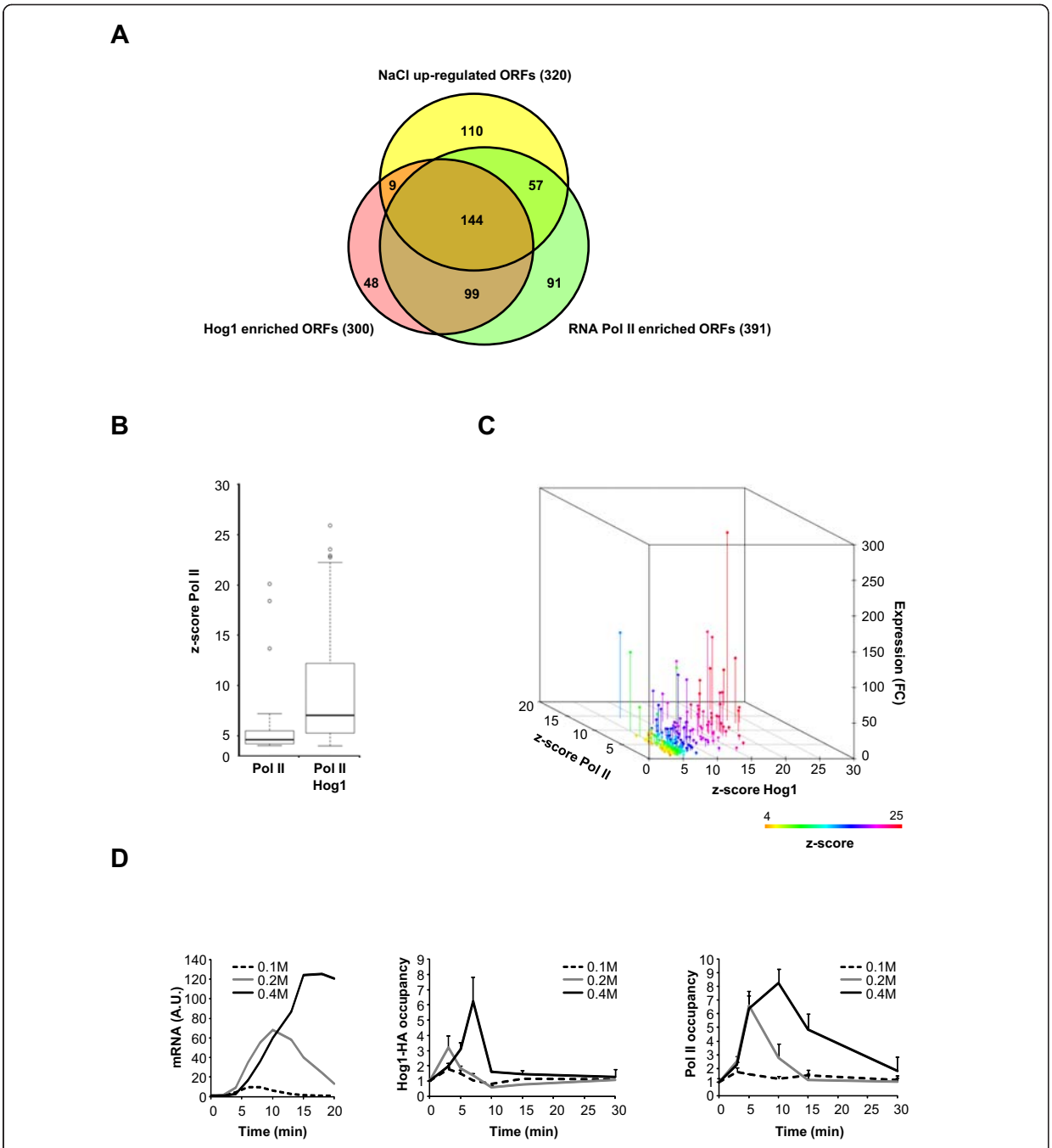


Figure 3 Greater enrichment of Hog1 induces stronger recruitment of RNA Pol II and determines transcriptional output. **(a)** A Venn diagram representing the overlap of Hog1-bound genes (*Z*-score > 4) and recruited RNA Pol II (*Z*-score > 6) with 320 representative genes that are up-regulated upon osmostress by more than three-fold. **(b)** Colocalization of Hog1 with RNA Pol II results in stronger binding of both to chromatin. The boxplot represents distribution of *Z*-score values of genes with Hog1 enrichment and RNA Pol II compared to those genes with RNA Pol II but without Hog1 association. **(c)** Positive correlation between Hog1, RNA Pol II binding and expression in response to stress. Each point in the three-dimensional graph represents the 144 genes (overlapping region from (a)); the x-axis represents binding of Hog1, the y-axis binding of Rbp1 (both as *Z*-score values) and the z-axis the expression as the fold change (*FC*) of a wild-type strain upon stress. **(d)** Gene expression is dependent on the duration and intensity of Hog1 binding to target genes. The dose-response expression of osmoresponsive genes was assayed by northern blot and probed for *STL1* (left-hand panel). Association of tagged Hog1-HA (middle panel) and Rpb1 (right-hand panel) to the promoter region of *STL1* was analyzed by ChIP. The results are shown as the fold induction of stressed against non-stressed (time zero) cultures.

recruitment of RNA Pol II and transcription. We followed *STL1* expression in response to 0.1, 0.2 and 0.4 M NaCl together with the association of Hog1 and RNA Pol II. Weak expression of *STL1* was observed at 0.1 and 0.2 M NaCl, in clear contrast to the induction observed with 0.4 M NaCl (similar results were observed for *CTT1* and *ALD3*). Remarkably, the initial recruitment of RNA Pol II at *STL1* was similar at 0.2 and 0.4 M NaCl; however, the residence time of Hog1 at the loci was clearly shorter (Figure 3d; Figure S8 in Additional file 1). Thus, the association time of Hog1 with stress loci appears to be crucial for determining the degree of gene induction upon osmostress.

Hog1 mediates chromatin changes at stress-responsive loci

Hog1 stimulates chromatin remodeling at specific stress-responsive loci by recruiting the RSC chromatin remodeler [21]. We investigated whether all stress-responsive genes were subjected to changes on chromatin organization and the relevance of the SAPK to those changes. We used genome-wide MNase digestion of chromatin and deep sequencing (MNase-Seq) before and after stress. Wild-type and *hog1* strains were subjected (or not) to osmostress and cells were fixed before digestion of chromatin by MNase to prevent Hog1 activation during the preparation of spheroplasts (see Materials and methods). The nucleosomal profile around the transcription start site (TSS) in genes that were not regulated upon stress did not change when cells were subjected to osmostress (Figure 4a, upper panel). We then analyzed nucleosome positioning in stress-induced genes with expression that does not depend on Hog1. Upon osmostress there were slight changes of nucleosome occupancy, especially around the TSS and those changes were similar in *hog1*-deficient cells (Figure 4a, middle panel). In clear contrast, when Hog1-dependent genes were analyzed, a dramatic change of nucleosome occupancy occurred upon stress at both the promoter and ORF regions. These changes on nucleosomes were completely abolished in *hog1* cells. It is noteworthy that the +1 nucleosome in stress-responsive genes appears to be shifted slightly when compared to localization of the genome-wide +1 nucleosome, suggesting a particular chromatin structure for stress-responsive genes. Color maps of -1,000 to +1,000 of each ORF aligned by TSS have also been included to illustrate nucleosome organization genome-wide and in stress-responsive genes (Figure S9 in Additional file 1). Taken together, efficient nucleosome re-organization at stress-responsive genes is completely dependent on the SAPK (Figure 4a, lower panel). When changes on the chromatin structure were quantified (percentage of nucleosome occupancy), we found that osmostress induces a 25% decrease in nucleosome occu-pancy in Hog1-independent genes (similar to that found in wild-type and *hog1* cells), whereas it was decreased 51% in Hog1-dependent genes (Figure 4b). Thus, Hog1 is crucial to inducing major changes in chromatin structure.

In summary, genome-wide binding studies in combination with analysis of chromatin structure have shown that Hog1 serves to bypass the general down-regulation of gene expression that occurs in response to stress. Hog1 permits efficient targeting of the RNA Pol II machinery and, furthermore, induces major changes of chromatin structure at stress-responsive loci. The combination of targeted recruitment of RNA Pol II with chromatin remodeling is essential to maximize gene expression in response to external stimuli.

Conclusions

In response to an environmental insult such as osmostress, there are major changes of RNA Pol II localization towards stress-responsive genes, in contrast to housekeeping genes, which correlates with down-regulation of general transcription. This indicates that while there is general down-regulation of expression and RNA Pol II association with housekeeping genes, RNA Pol II is recruited to stress-responsive genes. We showed that this RNA Pol II relocalization is a phenomenon that requires the Hog1 SAPK. We analyzed genome-wide localization of the SAPK and found its presence in Hog1-dependent genes more frequently than in any published study and, remarkably, we found Hog1 associated with RNA Pol III genes, showing for the first time the interaction of the SAPK with the RNA Pol III machinery. We undertook genome-wide analysis of the change of chromatin structure upon stress. We showed that there is strong remodeling of chromatin restricted to stress-responsive genes, which depends completely on the SAPK. Hog1 association and chromatin remodeling correlated very well with higher levels of transcription, which highlights the relevance of chromatin regulation and transcription.

Hog1 serves to bypass the general down-regulation of gene expression that occurs in response to osmostress by targeting RNA Pol II machinery and by inducing chromatin remodeling at stress-responsive loci.

Materials and methods

Yeast strains and plasmids

Saccharomyces cerevisiae strain BY4741 (MATa *his3-Δ1 leu2-Δ0 met15-Δ0 ura3-Δ0*) and its derivatives YGM61 (*HOG1*::*KanMX4*), the carboxy-terminal tagged strain YEN173 (Hog1-6HA::*HIS3*), YMN07 (Rpc82-18myc:: *KanMX4*), YMN10 (Rpc82-18myc:: *KanMX4 HOG1*:: *URA3*) and YMN14 (Rpc8218myc::*KanMX4* Hog1-6HA:: *HIS3*) were used. Epitope tagging or gene deletions were done with a PCR-based strategy.

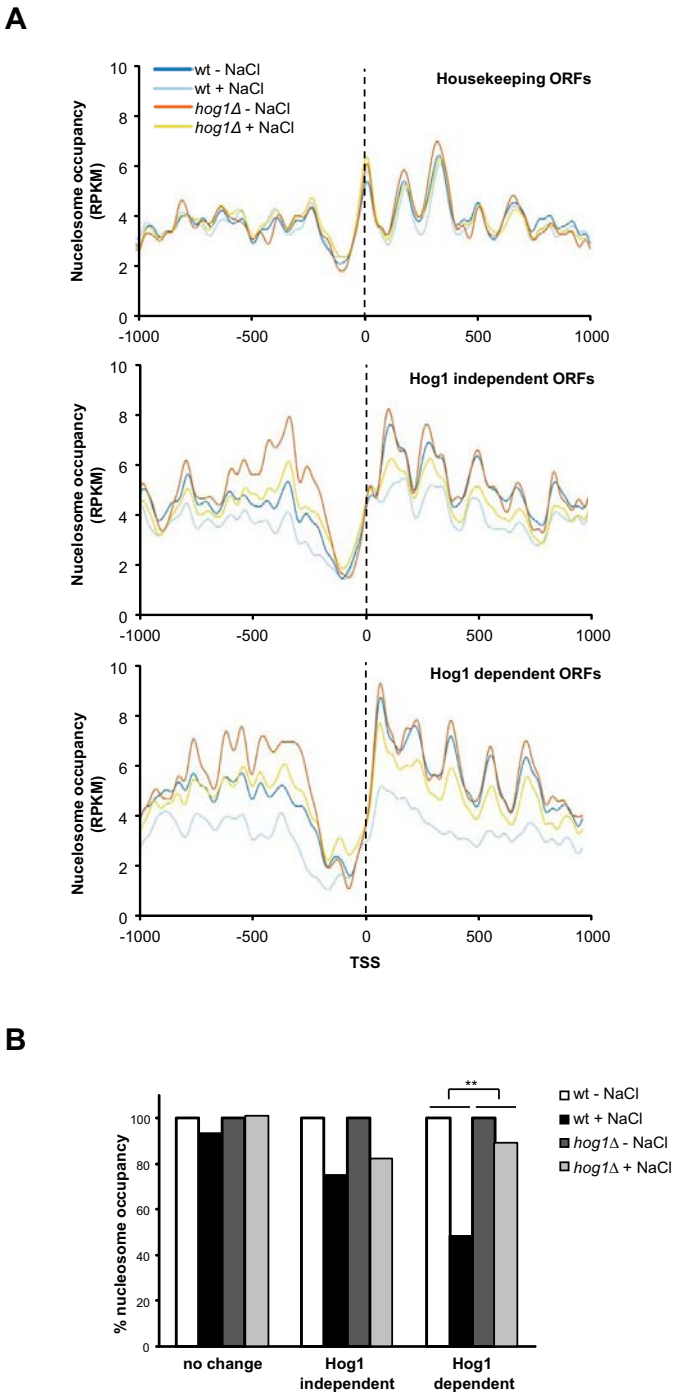


Figure 4 Hog1 mediates major changes in chromatin structure to facilitate transcription of stress-responsive genes. **(a)** The distribution of nucleosomes referenced as reads/kilobase per million mapped reads (RPKM; transcription start site (TSS) ± 1 kb) of wild-type (wt; blue) and *hog1* mutant (red) strains subjected (or not) to osmostress as indicated in the key. The plot represents the mean of reads in a group of 100 genes without transcription changes upon stress (upper panel), 100 genes that are induced in a Hog1-independent manner (middle panel) and 100 genes whose expression is induced upon osmostress depending on the SAPK (lower panel) (as in Figure 1d). The dotted black line marks the TSS. **(b)** The percentage of nucleosome occupancy in subsets from (a). Nucleosome occupancy was determined by averaging the TRPKs (trimmed mean of *M* values normalized read/kilobase density) from the 200 bp immediately downstream of the TSS. Average reads of non-stressed samples was used as maximum occupancy and as a reference for treated samples. **The statistical significance of the difference was assessed by a paired Student *t*-test of acceptance of equality at (*P*-value < 0.01) comparing the eviction of wild-type versus *hog1* upon stress.

ChIP-Seq

Wild-type and *hog1* mutant (YGM61) *S. cerevisiae* strains were grown to early log phase and exposed to osmostress (0.4 M NaCl) for 5 minutes for Hog1 immunoprecipitation (anti-HA 12CA5) and for 10 minutes for RNA Pol II immunoprecipitation (8WG16, Covance, Richmond, CA, USA). ChIP was done as described [10,24]; 10 ng of DNA from each ChIP and condition was used to create sequencing libraries and then subjected to 36-nucleotide single-read sequencing on a Solexa Genome Analyzer IIx instrument. Only sequencing reads that mapped to only one location were aligned to the *S. cerevisiae* genome (sacCer2) allowing up to three mismatches/reads. Chip reads were extended by 250 bp and normalized using Pyicos software [25]. Enrichment of Hog1 and RNA Pol II was done by running the Pyicos enrichment protocol [25] comparing untreated to treated samples. Hog1-dependent RNA Pol II recruitment was determined by comparing salt-treated samples with wild-type and *hog1* strains. For all comparisons, enrichment was considered significant for a *Z*-score > 4 ($P = 0.0001$). *MA* plots were done using the Pyicos software, where *M* represents the log ratio of stressed versus non-stressed (y-axis) and *A* is the average of the log intensities (x-axis) of all the genes, for enrichment of Hog1 and RNA Pol II. *MA* plots for Hog1-dependent RNA Pol II enrichment (Figure 1d, right-hand panels) compare NaCl-treated samples in a wild-type versus *hog1* strain. ChIP-Seq data have been deposited at the NCBI Gene Expression Omnibus (GEO) database with accession number GSE41494.

MNase nucleosome mapping

Wild-type and *hog1* mutant (YGM61) strains were grown to early log phase in YPD medium, then subjected (or not) to osmostress (0.4 M NaCl, 10 minutes). Spheroplasts and digestion with MNase were done essentially as described but with some modifications [21]. Spheroplasts were prepared from mid-log phase cultures grown in YPD medium, following crosslinking with 1% (v/v) formaldehyde for 20 minutes, treated with 125 mM glycine for 15 minutes and washed four times with Tris-buffered saline (20 mM Tris-HCl pH, NaCl 150 mM). Cells were then lysed and immediately digested with 7.5 to 125 mU of MNase (Worthington Biochemical Corporation, Lakewood; NJ., USA). DNA was subjected to electrophoresis in a 1.5% (w/v) agarose gel and then stained with ethidium bromide. The band corresponding to the mononucleosome was cut from the gel and purified using a QIAquick gel extraction kit (Qiagen, Chatsworth, CA, USA). Mononucleosomal DNA from each sample used to create sequencing libraries was subjected to 36-nucleotide single-read sequencing on a Solexa Genome Analyzer IIx. Sequencing reads were aligned to the *S. cerevisiae* reference genome using Bowtie software [26] allowing up to three mismatches/reads. Over-represented reads were eliminated to reduce PCR amplification artifacts. Coverage values were calculated for each position on the genome, normalized and converted to reads per million. The mean reads per million was obtained by aligning genes to the TSS. The TSS used was obtained from the 'saccer2' database available from the Saccharomyces Genome Database (2008). All the annotated features, including the TSS, come from the Saccharomyces Genome Database annotation. Peak detection was done with the *nucleR* software package [27]. The percentage of nucleosome occupancy was determined by running an extension and normalization protocol using the first 200 bp downstream of the TSS, which encompasses the +1 nucleosome for all the genes clusters, using Pyicos software [25]. Eviction was determined by setting the non-treated nucleosome occupancy in TRPK (trimmed mean of *M* values normalized read/kilobase density) at time zero of each strain as the maximum occupancy (100%) and used as the reference for treated samples. MNase-Seq data have been deposited at the NCBI GEO site with accession number GSE41494.

Chromatin immunoprecipitation

Yeast cultures were grown to early log phase, then subjected to osmostress at different salt concentrations (0.1 M, 0.2 M or 0.4 M NaCl); ChIP was done as described [13]. Antibodies used for immunoprecipitation were anti-HA 12CA5 for Hog1-HA, 8WG16 (Covance) for Rpb1 or anti-myc, 9E10 for Rpc82-Myc. For crosslinking, yeast cells were treated with 1% formaldehyde for 20 minutes at 25°C. Conventional and real-time PCR analysis of stressed and constitutively expressed genes used the following primers with locations indicated by the distance from the respective ATG initiation codon: *STL1* (promoter, -372/-112; ORF, 1,475/1,575), *CTT1* (promoter, -432/-302; ORF, 736/836), *PMA1* (+1,010/+1,250), *SCR1* (+19/+369), *RPR1* (+100/+325), tP(UGG)O3 (-61/+185) and tF(GAA)D (-74/+226) and *TEL* (telomeric region on the right arm of chromosome VI). Experiments were done with three independent chromatin preparations and quantitative PCR analysis was done in real time with a sequence detector (ABI 7700, Applied Biosystems, Foster City, CA., USA). Immunoprecipitation efficiency was calculated in triplicate by dividing the amount of PCR product in the immunoprecipitated sample by that in the *TEL* sequence control. The binding data are presented as fold induction with respect to the non-treated condition.

Northern blot analysis

Yeast cultures were grown to early log phase (absorbance at 660 nm 0.6 to 0.8). Cells were subjected (or not) to 0.1 M, 0.2 M or 0.4 M NaCl. Total RNA was probed by using radiolabeled fragments *ALD3* (1.5 kb), *STL1* (1.7 kb), *CTT1* (1.7 kb) and *ENO1* (1.3 kb). Signals were quantified

with a phosphorimager (Typhoon 8600, Molecular Dynamics, Sunnyvale, CA., USA) using ImageQuant software (Molecular Dynamics, Sunnyvale, CA., USA).

In vivo co-precipitation assay

Rpc82 and/or Hog1-tagged cells in mid-log phase were treated (or not) with 0.4 M NaCl for 15 minutes and then collected extracts were subjected to immunoprecipitation as described [20] with anti-HA 12CA5 or anti-Myc 9E10. Proteins were detected by western blotting against Hog1 (anti-Hog1, Santa Cruz Biotechnology, Santa Cruz, CA,, USA) and anti-Myc 9E10.

Gene expression studies

Microarray experiments and data analysis were done as described [28]. Briefly, three independent cultures of wild-type strains were grown to exponential phase in YPD medium and stressed (or not) with 0.4 M NaCl for 15 minutes. Gene expression was determined by comparing the expression of the non-stressed versus stressed conditions. RNA microarray data have been deposited at the NCBI GEO with accession number GSE41451. On the bases of our microarray data together with the results of earlier studies [3], we have created different gene clusters depending on their expression under osmostress and their dependence on the presence of Hog1. The 'osmoresponsive genes' are a group of 662 ORFs whose fold change (FC) upon stress is > 1.75 in a wild-type stain (microarray data from this study), and the 'NaCl group' contains 320 genes with FC > 3 (Figure 1a and 2a, respectively). From these mentioned groups, we chose a subset of 100 genes whose expression is considered Hog1-dependent or Hog1-independent with the following criteria: 'Hog1-dependent genes' are those genes whose expression depends at least 25% on the presence of the mitogen-activated protein kinase (MAPK) described in [3]. 'Hog1-independent genes' are those genes whose expression in a *hog1* strain is at least 90% similar to the wild-type strain described in [3]; moreover, genes that showed enrichment for Hog1 at their coding regions but their expression was not considered Hog1-dependent have been considered independent. The 'housekeeping genes' are those genes whose expression upon osmostress remained unchanged (FC 1 to 1.1). The 'down-regulated genes' are those genes whose expression was at least two-fold lower upon osmostress. See Additional file S2 for the complete list of genes included in each category.

Additional material

Additional file 1: Supplementary Figure S1 to S9.

Additional file 2: Supplemetary Table 1. List of all the genes considered in the manuscript as Hog1-dependent (top 100 Hog1-depedent osmoresponsive genes), Hog1-independent (top 100 Hog1-independent osmoresponsive genes) and a list of all osmoresponsive genes from the microarray analysis (see the 'Gene expression studies' section in Materials and methods for the criteria used to define the genes present in the list).

Abbreviations

ChIP: chromatin immunoprecipitation; FC: fold change; GEO: Gene Expression Omnibus; HOG: high osmolarity glycerol; MNase: micrococcal nuclease; ORF: open reading frame; Pol: polymerase; SAPK: stress-activated protein kinase; TSS: transcription start site.

Acknowledgements

The authors thank L Subirana, S Ovejas and A Fernandez for technical support and C Solé (UPF), Sebastián Chávez and Gonzalo Millán-Zambrano (Universidad de Sevilla) for experimental support and helpful comments. MN is a recipient of an FIS fellowship. This work was supported by grants from the Spanish Government (BIO2011-23920 to EE, BIO2009-07762 and BFU2012-33503 to FP, BFU2011-26722 to EN), the Consolider Ingenio 2010 programme CSD2007-0015 to FP and FP7 UNICELLSYS grant 201142, the Fundación Marcelino Botín (FMB) to FP. EN and FP are recipients of an ICREA Acadèmia (Generalitat de Catalunya).

Author details

[1]Cell Signaling Unit, Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra (UPF), E-08003 Barcelona, Spain. [2]Joint IRB-BSC Program on Computational Biology, Institute for Research in Biomedicine, Baldiri i Reixac 10, 08028 Barcelona, Spain. [3]Computational Genomics, Universitat Pompeu Fabra, Dr Aiguader 88, E08003 Barcelona, Spain. [4]Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluís Companys 23, E08010 Barcelona, Spain.

Authors' contributions

MN and NC did most of the experiments and analysis. OF, MO, JG and EE did the bioinformatics analysis. MN, NC, EN and FP did the experimental designs and wrote the paper. All authors have read and approved the manuscript for publication.

Competing interests

The authors declare that they have no competing interests.

Received: 25 June 2012 Revised: 29 October 2012
Accepted: 18 November 2012 Published: 18 November 2012

References

1. de Nadal E, Ammerer G, Posas F: **Controlling gene expression in response to stress.** *Nat Rev Genet* 2011, **12**:833-845.
2. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO: **Genomic expression programs in the response of yeast cells to environmental changes.** *Mol Biol Cell* 2000, **11**:4241-4257.
3. Posas F, Chambers JR, Heyman JA, Hoeffler JP, de Nadal E, Arino J: **The transcriptional response of yeast to saline stress.** *J Biol Chem* 2000, **275**:17249-17255.
4. Causton HC, Ren B, Koh SS, Harbison CT, Kanin E, Jennings EG, Lee TI, True HL, Lander ES, Young RA: **Remodeling of yeast genome expression in response to environmental changes.** *Mol Biol Cell* 2001, **12**:323-337.
5. Capaldi AP, Kaplan T, Liu Y, Habib N, Regev A, Friedman N, O'Shea EK: **Structure and function of a transcriptional network activated by the MAPK Hog1.** *Nat Genet* 2008, **40**:1300-1306.
6. Molin C, Jauhiainen A, Warringer J, Nerman O, Sunnerhagen P: **mRNA stability changes precede changes in steady-state mRNA amounts during hyperosmotic stress.** *RNA* 2009, **15**:600-614.
7. Ni L, Bruce C, Hart C, Leigh-Bell J, Gelperin D, Umansky L, Gerstein MB, Snyder M: **Dynamic and complex transcription factor binding during an inducible response in yeast.** *Genes Dev* 2009, **23**:1351-1363.
8. Romero-Santacreu L, Moreno J, Perez-Ortin JE, Alepuz P: **Specific and global regulation of mRNA stability during osmotic stress in Saccharomyces cerevisiae.** *RNA* 2009, **15**:1110-1120.
9. Miller C, Schwalb B, Maier K, Schulz D, Dümcke S, Zacher B, Mayer A, Sydow J, Marcinowski L, Dölken L, Martin DE, Tresch A, Cramer P: **Dynamic**

transcriptome analysis measures rates of mRNA synthesis and decay in yeast. *Mol Syst Biol* 2011, **7**:458.
10. Alepuz PM, Jovanovic A, Reiser V, Ammerer G: **Stress-induced map kinase Hog1 is part of transcription activation complexes.** *Mol Cell* 2001, **7**:767-777.
11. Proft M, Pascual-Ahuir A, de Nadal E, Arino J, Serrano R, Posas F: **Regulation of the Sko1 transcriptional repressor by the Hog1 MAP kinase in response to osmotic stress.** *EMBO J* 2001, **20**:1123-1133.
12. Proft M, Struhl K: **Hog1 kinase converts the Sko1-Cyc8-Tup1 repressor complex into an activator that recruits SAGA and SWI/SNF in response to osmotic stress.** *Mol Cell* 2002, **9**:1307-1317.
13. Alepuz PM, de Nadal E, Zapater M, Ammerer G, Posas F: **Osmostress-induced transcription by Hot1 depends on a Hog1-mediated recruitment of the RNA Pol II.** *EMBO J* 2003, **22**:2433-2442.
14. de Nadal E, Casadome L, Posas F: **Targeting the MEF2-like transcription factor Smp1 by the stress-activated Hog1 mitogen-activated protein kinase.** *Mol Cell Biol* 2003, **23**:229-237.
15. Pascual-Ahuir A, Struhl K, Proft M: **Genome-wide location analysis of the stress-activated MAP kinase Hog1 in yeast.** *Methods* 2006, **40**:272-278.
16. Pokholok DK, Zeitlinger J, Hannett NM, Reynolds DB, Young RA: **Activated signal transduction kinases frequently occupy target genes.** *Science* 2006, **313**:533-536.
17. Proft M, Mas G, de Nadal E, Vendrell A, Noriega N, Struhl K, Posas F: **The stress-activated Hog1 kinase is a selective transcriptional elongation factor for genes responding to osmotic stress.** *Mol Cell* 2006, **23**:241-250.
18. de Nadal E, Zapater M, Alepuz PM, Sumoy L, Mas G, Posas F: **The MAPK Hog1 recruits Rpd3 histone deacetylase to activate osmoresponsive genes.** *Nature* 2004, **427**:370-374.
19. Zapater M, Sohrmann M, Peter M, Posas F, de Nadal E: **Selective requirement for SAGA in Hog1-mediated gene expression depending on the severity of the external osmostress conditions.** *Mol Cell Biol* 2007, **27**:3900-3910.
20. Sole C, Nadal-Ribelles M, Kraft C, Peter M, Posas F, de Nadal E: **Control of Ubp3 ubiquitin protease activity by the Hog1 SAPK modulates transcription upon osmostress.** *EMBO J* 2011, **30**:3274-3284.
21. Mas G, de Nadal E, Dechant R, Rodríguez de la Concepción ML, Logie C, Jimeno-González S, Chávez S, Ammerer G, Posas F: **Recruitment of a chromatin remodelling complex by the Hog1 MAP kinase to stress genes.** *EMBO J* 2009, **28**:326-336.
22. Proft M, Struhl K: **MAP kinase-mediated stress relief that precedes and regulates the timing of transcriptional induction.** *Cell* 2004, **118**:351-361.
23. Pelet S, Rudolf F, Nadal-Ribelles M, de Nadal E, Posas F, Peter M: **Transient activation of the HOG MAPK pathway regulates bimodal gene expression.** *Science* 2011, **332**:732-735.
24. Kuras L, Struhl K: **Binding of TBP to promoters in vivo is stimulated by activators and requires Pol II holoenzyme.** *Nature* 1999, **399**:609-613.
25. Althammer S, Gonzalez-Vallinas J, Ballare C, Beato M, Eyras E: **Pyicos: a versatile toolkit for the analysis of high-throughput sequencing data.** *Bioinformatics* 2011, **27**:3333-3340.
26. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol* 2009, **10**:R25.
27. Flores O, Orozco M: **nucleR: a package for non-parametric nucleosome positioning.** *Bioinformatics* 2011, **27**:2149-2150.
28. Ruiz-Roig C, Vieitez C, Posas F, de Nadal E: **The Rpd3L HDAC complex is essential for the heat stress response in yeast.** *Mol Microbiol* 2010, **76**:1049-1062.

doi:10.1186/gb-2012-13-11-r106
Cite this article as: Nadal-Ribelles *et al.*: Hog1 bypasses stress-mediated down-regulation of transcription by RNA polymerase II redistribution and chromatin remodeling. *Genome Biology* 2012 **13**:R106.

2.4. Modeling genome-wide kinetics of endo/exonuclease digestion

I would like to mention this work in the results section of this thesis for three reasons. The first one is that it implied a short-stay in the group Prof. Jason Lieb, former Director of the Carolina Center for Genome Sciences and current member of the Lewis-Sigler Institute for Integrative Genomics of the Princeton University. Of course, this is a top group in genomics with a large and acknowledged trajectory, supported by many outstanding publications. Second, this collaboration was an opportunity for me to spend the summer of 2013 in the University of North Carolina in Chapel Hill (USA), which of course was a very enriching experience, both in a professional and personal point of view. Third reason is, thanks to this collaboration, we could benchmark the methods and the expertise obtained in other works presented in this thesis in a completely different environment – worth to say, with quite good performance.

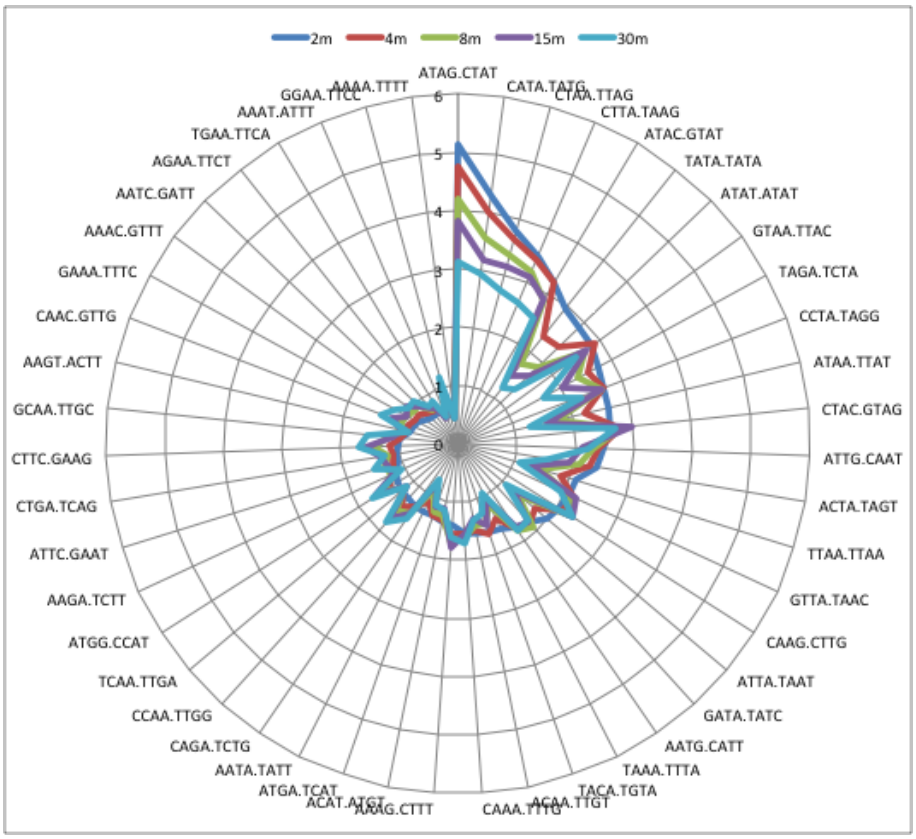


Fig. 27. MNase cutting site affinity for different tetramers. Radius represents the ratio of the cutting sites over the genome-wide expected frequency. Affinity is stronger in milder digestions but vanish as the digestion time increase.

During my stay in Chapel Hill, I worked in one of the projects of Tess Jeffers, a PhD student under the supervision of Prof. Lieb. The focus of this study was the kinetics of MNase digestion in *C.elegans,* using a time-course experiment. Five biological replicas with different digestion times (2', 4', 8', 15' and 30') enable the study of both the endo- and exo-nuclease activity of MNase under different degrees of saturation. Worth to say that it was extremely nice to collaborate in this project as it was a kind of wrap-up of many of the concepts we saw in our own work previously (MNase digestion patterns (page 35), time-course experiments (page 74) and the effects of under-/over-digestion of MNase (page 78)). In fact, the results obtained here validated some of our already published results, as the preferential patterns of MNase endo-nuclease (cleavage) activity (Fig. 27), and added a new dimension to the analysis by accounting for differential digestion times.

Combining some of the tools featured in the section *Algorithmic and computational methods*, we were able to detect and quantify the sensitivity of different nucleosomes to the MNase digestion according to their aggregated read length. This allowed the classification of regions with accumulations of long (resistant to digestion), short (especially sensitive to digestion) and dynamic (read length linked to the digestion time) (Fig. 28).

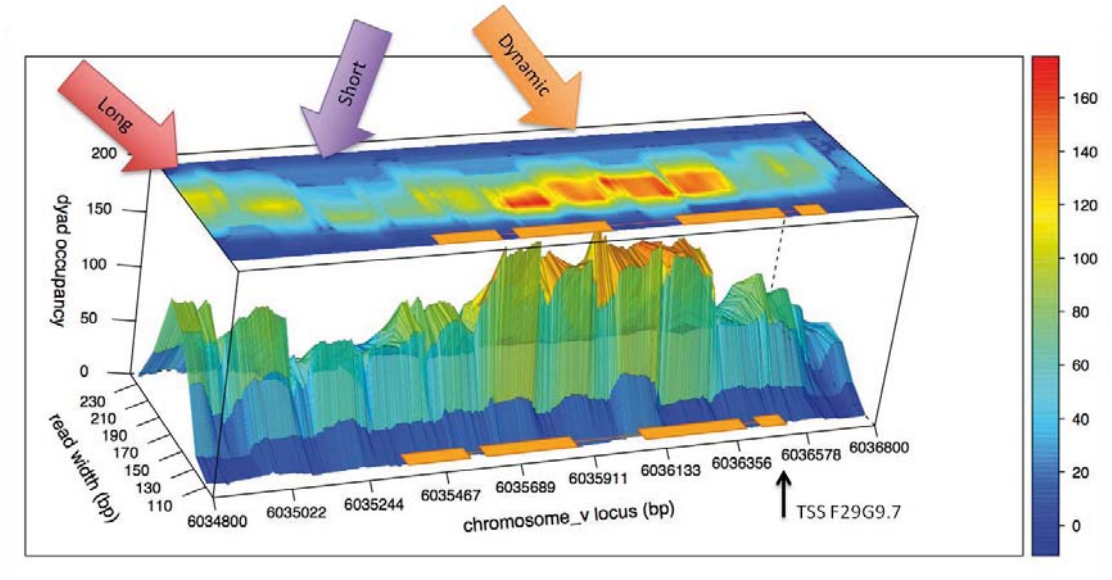


Fig. 28. 3D coverage plot around F29G9.7 gene body in C.Elegans. X axis shows the genomic locus, Y axis the read width and Z axis the occupancy of reads in a given position with a given width. A reduced representation in 2D is shown in the top of the box in an electrophoresis gel like style. Arrows point to regions with accumulations of long, short and dynamic reads, as described above.

Finally, we classified regions that were covered in short-digestions but not in longer ones as *unstable* regions. The study of physical properties of DNA in these unstable regions reveals an intrinsically high local stiffness (Fig. 29). This reinforces the conclusions regarding the MNase sensitivity in specific regions of the genome with special mechanistic properties we found in our article *Physical properties of naked DNA influence nucleosome positioning and correlate with transcription start and termination sites in yeast* (page 35).

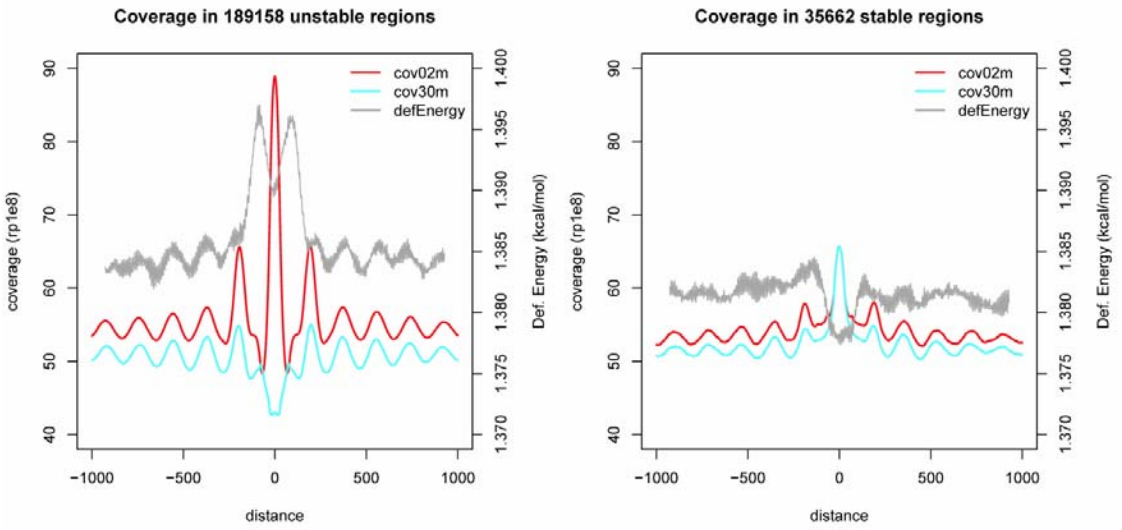


Fig. 29. Coverage in regions classified as unstable (with different coverage patterns in the 2 minute MNase digestion sample and the one digested during 30 minutes) and stable regions (with the same patterns in both experiments). Energy profile (grey line) shows a higher mean value in unstable regions than in stable ones. Furthermore, nucleosome occupancy peaks are positioned around those regions following the energetic barrier model described previously (page 79)

In summary, our previous expertise in the different topics covered here allowed an efficient collaboration in a short period of time. C.elengans and S.cerevisiae share common features regarding the role of DNA flexibility and enzyme dynamics. The tools and approaches that we used in the study of yeast chromatin can be also applied without problem in larger organisms. Lastly, maybe a relevant point we can highlight from this collaboration is the fact that, despite we are a relatively new group in genomics, we can offer valuable contributions to more expert top-groups in the field.

3. Algorithmic and computational methods

Due to my background in computer science, the development of new methods and approaches for analysis is always challenging and motivating. In fact, my first approach to research was the development of a probabilistic bacteria identifier in collaboration with the microbiology department of the University of Barcelona (Flores *et al.*, 2009). The possibility to do a PhD in the Joint Program for Computational Biology between the Institute for Research in Biomedicine and the Barcelona Supercomputing Center was for me an attractive offer where both massive experimental data and computational resources will be available. As presented below, I took advantage of this position by creating new methods for the computational analysis of chromatin which we made available to the scientific community.

All the tools presented here are based on the R/Bioconductor framework (Gentleman *et al.*, 2004) and somehow related to the integrative analysis of high-throughput data. R/Bioconductor has become the election of many bioinformaticians due to the flexibility and power of this array-based language and the vast repository of tools it features. Among these tools, we will find libraries for preprocess, analyze and visualize biological data including sequences, transcription, genomic features and protein structures.

The Bioconductor community is one of the most active user's groups in bioinformatics. It provides a valuable source of knowledge where almost any issue can be solved. Another distinctive trait of Bionconductor is that all packages include a mandatory user-level, step-by-step guide to the different functions of different packages in a document called *vignette*. All submitted packages to the Bioconductor repository are peer-reviewed in terms of efficiency, redundancy and functionality and only after a critical assessment their publication is allowed.

In the next pages, we will present four different libraries we developed during this PhD thesis:

- *nucleR: a package for non-parametric nucleosome positioning* **(page 109)**
- *htSeqTools: high-throughput sequencing quality control, processing and visualization in R* **(page 115)**
- *DASiR: Programmatic data retrieval in R* **(page 118)**
- *Nucleosome Dynamics: Comparative analysis of nucleosome positioning at read level* **(page 125)**

3.1. nucleR: a package for non-parametric nucleosome positioning

nucleR was my first publication in MMB. This library was the result of packing together some of the scripts I used in the processing of our first nucleosome maps. As seen in the introduction (page 20), a common step in the processing of nucleosomal data is to convert the experimental occupancies into an array of regions each one representing a single, averaged, nucleosome position. We refer to this process as nucleosome calling.

In the first genome-wide nucleosome maps published (Yuan *et al.*, 2005; Lee *et al.*, 2007) the low resolution of the nucleosome occupancy signal, usually obtained using tiling arrays, required the use of statistical models to obtain the nucleosome calls. In a first approach, I tried to reproduce a Hidden Markov Model proposed previously (Yuan *et al.*, 2005) but, as stated in Introduction, this approach presents some important flaws (page 20).

The main problems found in the preprocessing of nucleosome maps were two: the presence of missing areas (i.e., low coverage areas) and the noisy and fuzzy nature of the nucleosome occupancy signals. HMMs allow dealing with this kind of problems, revealing then the "hidden" nucleosome patterns below a noisy or incomplete signal, but their application in high resolution data was really inefficient.

nucleR solves the first problem, areas with missing information, by interpolating missing bases from adjacent points. This allows the conversion of a tiling array signal where information is spanned every 20bp to 1bp-resolution data, comparable to sequencing experiments. For the second problem, the large presence of noise in the nucleosome coverage signal, nucleR uses a two-step approach. In the first step, only required for sequencing data, the signal-to-noise ratio is increased by working only with the central bases of the sequencing reads instead of the whole 147 bases. This can be achieved by shifting the reads in single-end sequencing or trimming them in the case of paired-end sequencing. As a side effect, this allows the correction of reads mapping in opposite strands for the same putative nucleosome (Fig. 30).

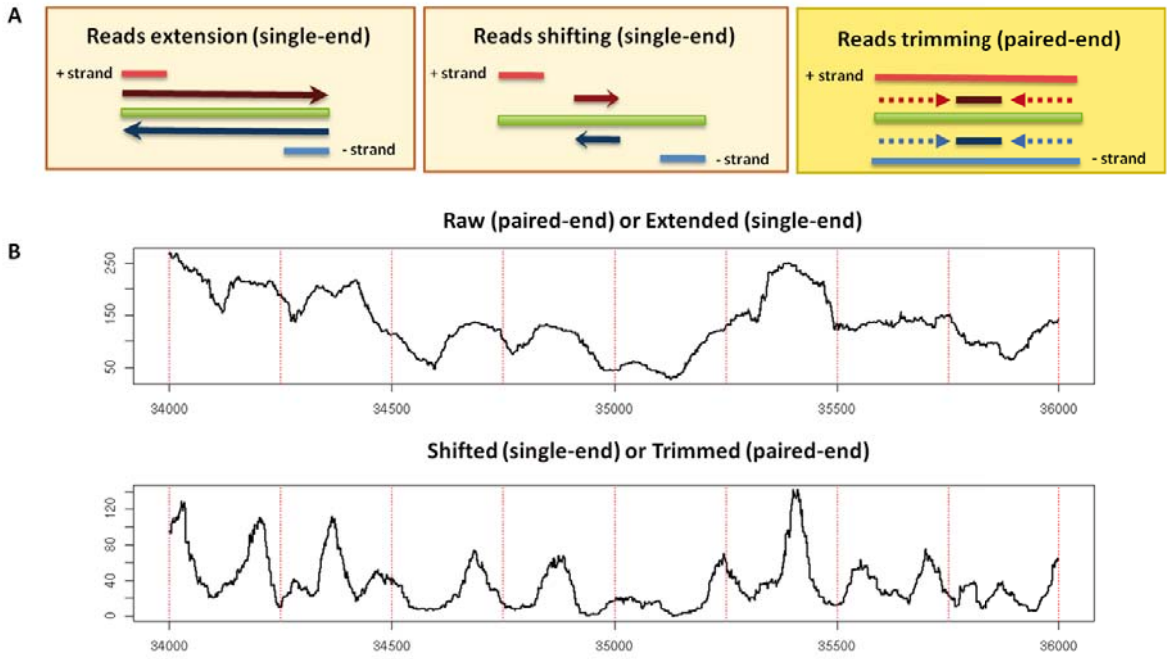


Fig. 30 Read-level pre-processing in high-throughput sequencing. A) In single-end sequencing, obtained reads from different strands (bright red, + strand; bright blue, - strand) can be extended (left) or shifted (right) downstream to match the expected fragment width (147bp in a case of a nucleosome, in green) to obtain corrected reads (dark red, dark blue). In the case of paired-end reads (right), reads can be trimmed to remark the dyad position. B) The same region with different processing. Bottom plot, corresponding to shifted or trimmed reads remarks the dyad positions, increasing the signal-to-noise ratio and allowing a clearer detection of nucleosomes.

After this process, the next step is the detection of occupancy peaks. As can be observed in the bottom panel of Fig. 30, the accumulations of reads around a given position appear as coverage peaks easily visible to naked eye. The challenge here was how to perform the peak calling process in a still noisy coverage where a simple threshold will detect many adjacent peaks due to the irregularity of the profile. This problem was bigger in the case of tiling array data, which, after the interpolation, presented a huge amount of noise (Fig. 31, left panel). Smoothing the coverage by means of a running window was not a good option, as the averaged profile could merge adjacent peaks(Fig. 31).

The differential approach of *nucleR* respect other existing methods available on that time is that, instead of performing a smoothing in the coverage to decrease the noise, the coverage signal is filtered with a noise gate. This is a common procedure in telecommunications and can be implemented by analyzing the power-spectrum of the signal in the Fourier space.

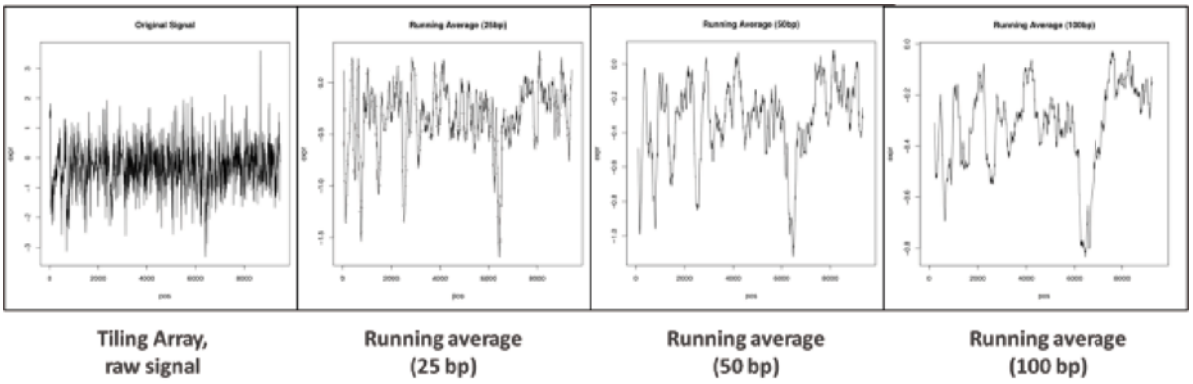


Fig. 31. The raw tiling array signal (left-most plot) in a random locus is subsequently smoothed by using and increasing window-width running average. Smoother profiles allow a better identification of peaks, but peaks corresponding to adjacent nucleosomes can be merged in large windows.

The differential approach of *nucleR* respect other existing methods available on that time was, instead of just smoothing the noise, filter the coverage signal with a noise gate. This is a common procedure in telecommunications and can be implemented by analyzing the power-spectrum of the signal in the Fourier space. Using the Fast Fourier Transform (FFT), nucleR converts the input coverage into a power-spectrum (the real component of the FFT) where different frequencies are weighted individually. Then, a certain number of components (around 1-2% of the input length depending of the source of the data) are selected, meanwhile frequencies attributed to noise and echoes of lower-frequencies, are knock-out. With the inverse FFT, the original signal is reconstructed but without noise (Fig. 32).

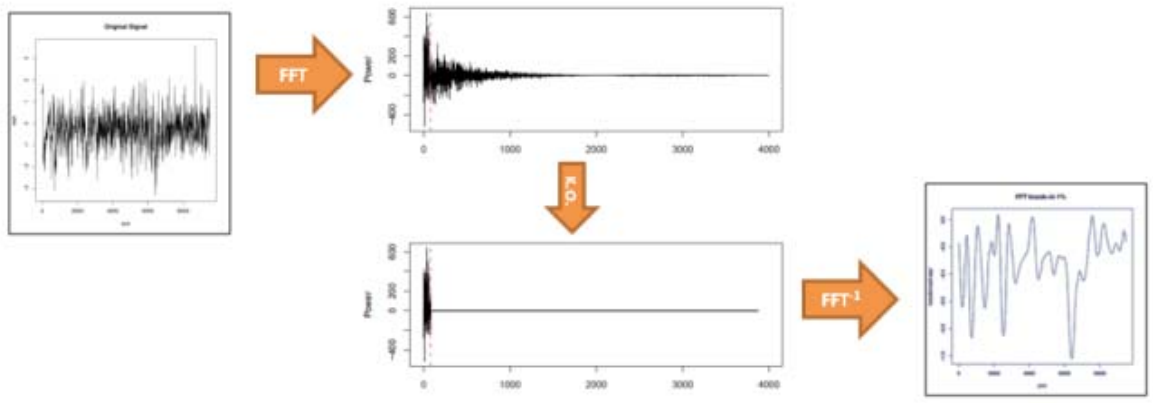


Fig. 32 Schema of noise filtering. Input noisy signal is converted to Fourier Space and components over a dynamic threshold are set to 0 (knock-out). After the inverse FFT, we obtain a profile without noise.

After the processing and filtering, the detection of local maxima search can be implemented just by looking changes in the signal trend (computational cost O(n)). Worth to say that the filtering process with FFT uses some algorithmic optimizations which ensure a fixed computational cost of O(n·log(n)) and lineal cost in memory. After the detection of peak summits, *nucleR* scores the fuzziness and the coverage of every peak, providing a measure of the goodness of positioning of every call (Fig. 33).

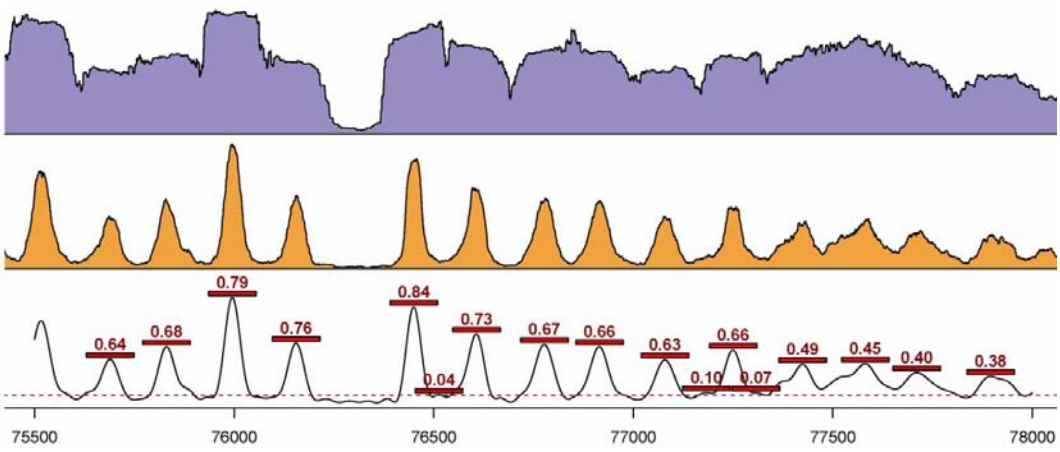


Fig. 33. Example of nucleR application over a paired-end sequencing experiment. Full length reads (orange) are trimmed to remark the dyad (purple). Small fluctuations over this dyad occupancy are filtered with the FFT (cyan profile) and peak summits are annotated as nucleosome calls (red boxes).

nucleR's approach was simpler, faster and more accurate than any other nucleosome calling method available upon its publication. Since its publication in June of 2011 and till March 2014, it has been downloaded more than 6500 times from Bioconductor repository and it is still a reference nucleosome caller featured in more than 17 publications (data from March-2014).

Publication:

Flores, O. and Orozco, M. (2011) "*nucleR: a package for non-parametric nucleosome positioning*", Bioinformatics (Oxford, England) 27: 2149–50.

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

BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 15 2011, pages 2149–2150
doi:10.1093/bioinformatics/btr345

Genome analysis

Advance Access publication on June 7, 2011

nucleR: a package for non-parametric nucleosome positioning

Oscar Flores[1] and Modesto Orozco[1,2,3,*]

[1]IRB-BSC Joint Research Program on Computational Biology, Institute of Research in Biomedicine, Baldiri i Reixac 10, [2]Department of Biochemistry and Molecular Biology, University of Barcelona, Avinguda Diagonal 645 and [3]Instituto Nacional de Bioinformática. Parc Científic de Barcelona. Baldiri i Reixac 10, Barcelona 08028, Spain

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: nucleR is an R/Bioconductor package for a flexible and fast recognition of nucleosome positioning from next generation sequencing and tiling arrays experiments. The software is integrated with standard high-throughput genomics R packages and allows for *in situ* visualization as well as to export results to common genome browser formats.

Availability: Additional information and methodological details can be found at http://mmb.pcb.ub.es/nucleR

Contact: modesto.orozco@irbbarcelona.org

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on March 15, 2011; revised on April 26, 2011; accepted on June 1, 2011

1 INTRODUCTION

Eukaryotic chromatin is organized in nucleosomes, a structure with approximately 147 basepairs (bp) of DNA wrapped around an octamer of histones (Jiang and Pugh, 2009). Nucleosomes affect the packaging and accessibility of DNA, thus playing a crucial role in defining its functionality (Jiang and Pugh, 2009). The development of high-throughput techniques, such as tiling arrays (TA) and next generation sequencing (NGS), coupled to Micrococal Nuclease (MNase) digestions has enabled the study of nucleosome positioning at the entire genome level for several organisms including humans (Jiang and Pugh, 2009). Some clear features emerged from these studies, such as the presence of well-positioned nucleosomes and their depletion in regions surrounding transcription start sites (TSS; Jiang and Pugh, 2009). However, well-positioned (phased across different cells) nucleosomes coexist with fuzzy (not-phased) ones outside the TSS. This variability makes nucleosome-positioning complex, requiring therefore the development of algorithms to find the 'most-probable' nucleosomal configuration. These approaches include, among others, Hidden Markov Models (HMM) (Lee *et al*., 2007; Yassour *et al*., 2008; Yuan *et al*., 2005), higher order Bayesian Networks (Chen *et al*., 2010) or mixed methods (Di Gesù *et al*., 2009).These methods are very powerful, but the intrinsic assumptions and the level of expertise of the modeler can significantly affect the results.

Here we present a new tool, nucleR, integrated in the open source, multiplatform R/Bioconductor framework. The approach is based on a fast, nonparametric detection of all nucleosome dyads and scoring of the calls. A good performance is achieved by filtering the noise using Fast Fourier Transform (FFT). The user has full freedom to export, select, merge or process suggested nucleosome calls in any desired way, making the method completely flexible. Algorithms presented here are suitable for most TA and single or paired ended NGS platforms.

*To whom correspondence should be addressed.

2 METHODS

nucleR's workflow is presented in Figure 1a. It relies the low-level processing of the genomic data to specialized R/Bioconductor packages, allowing for a wide variety of input formats.

The first step is to convert the input data to obtain 1bp resolution hybridization fluorescence ratios (TA) or short reads coverage (NGS). In the latter case, some extra manipulations are applied to the reads, like correcting the strand bias if working with single-ended sequencing or trimming the reads for remarking the position of the dyad in paired-end cases. Additionally, to reduce any potential bias due to the sequence preferences of MNase (Deniz *et al*., in press), coverage maps obtained from nucleosomal DNA can be easily corrected with those obtained in a parallel experiment for naked DNA. For TA, the main problem is the existence of DNA segments not covered by a probe, which in our procedure are inferred from neighboring probes.

The next step is 'profile cleaning' based on Fourier analysis, which simultaneously smoothes the signal and cleans the distortions in the coverage peaks. Noise removal from coverage profile is performed following signal theory (Smith, 1999). Accordingly, the original complex signal is described as a combination of simple periodic waves. By transforming the original profile into the Fourier Space using FFT, one can analyze the power spectrum of single frequencies, i.e. the contribution of every frequency to the original signal. High frequencies are usually echoes of lower frequencies and are sources of noise. They can therefore be removed without affecting the final profile (Smith, 1999). In our case, a small number of components are chosen depending on the nature of the experiment (typically 1% for TA and 2% for NGS; see Supplementary Material) and the rest are knockedout before performing the inverse FFT; see Figure 1b for an example of raw and filtered profiles. The following step is the detection of nucleosome dyad, which is done using a simple local maxima search and is largely facilitated by the clarity of the filtered profile. Nucleosome calls are determined by selecting the surrounding bases around the dyad position, and are scored based on the height and sharpness of the peak; giving high score to large and sharp peaks and penalizing fuzziness (Fig. 1b).

Once nucleosome calling and scoring is done, the user can manipulate the calls with standard R/Bioconductor tools to select, merge or perform further study on nucleosome positioning in a way that fulfills his/her specific needs. Methods for visualizing the results and exporting the data in BED and WIG formats are also provided. The nucleR package has been created to manipulate large datasets and offers an efficient usage of FFT (see Supplementary Material) and support for parallel processing in multicore machines.

3 RESULTS

In order to illustrate the performance of nucleR, we have analyzed two datasets derived from MNase treatment of yeast chromatin: TA

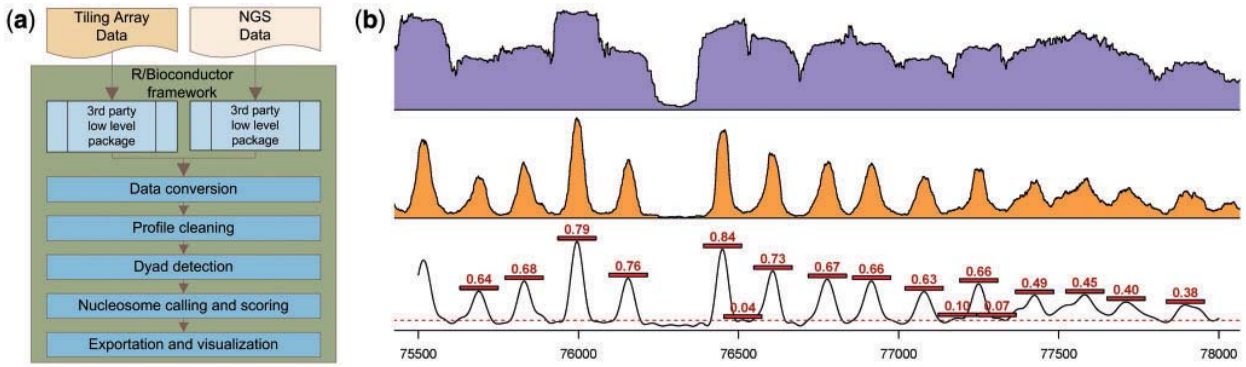


Fig. 1. (**a**) nucleR workflow diagram. (**b**) Top: raw coverage from NGS paired-end reads; middle: coverage using trimmed reads, remarking the dyad location; bottom: filtered trimmed coverage and scored nucleosome positions. Detection threshold is marked as a red dashed line.

from Nislow's group (Lee *et al*., 2007) and an NGS experiment performed in our group (Deniz *et al*., in press). A comparative analysis has been performed using HMM results provided by Lee *et al*. (2007) and the package ChIPSeqR, available from R/Bioconductor repository. We selected the first method as it is widely used in the literature and the second one for being the only package that enables a nucleosome positioning analysis in R/Bioconductor.

The main difference between Lee's HMM and our approach is that the method presented here is able to detect multiple shifted nucleosomes in a given single position, and not just providing the 'most probable' state for each position. This has a large impact on the final map, where we are able to identify a richer landscape of nucleosome calls in TA experiments. Additionally, the HMM-based approach is difficult to apply to 1-bp resolution experiments, such as NSG, due to the large amount of memory required for backtracking. Apart from this scalability problem, the modeling of background transition probabilities, a key element in HMM, requires a very fine and subjective tuning. This can overconstrain the results, forcing, for example, preferred length linker DNA, strict periodic positioning or inability to detect coverage peaks due to strange chromatin structures, like centromers or tetrasomes. In the previous work of Lee *et al*., 70 873 nucleosomes (40 096 well positioned and 30 777 fuzzy) were detected. nucleR applied to the same dataset is able to detect a total of 151 882 individual scored nucleosome calls. Furthermore, our method displayed a larger ability to detect nucleosome positions in synthetic data (77% hits for nucleR versus 67% for HMM) and also a higher correlation when rebuilding the original read maps obtained from the TA experiment ($P=0.63$ for nucleR versus $P=0.38$ for HMM) (see Supplementary Material).

ChIPSeqR (www.bioconductor.org) provides a similar approach as the one presented here, but lacks support for TA data and a method for noise removal. This generates problems in peak detection since only global maxima above a settled threshold are detected efficiently, missing many relevant sub-peaks (local maximums) and leading to an underestimation of nucleosome density. In our NGS data nucleR detects 100 335 nucleosome calls (repeated genomic regions were not considered for this analysis), comprising all of the local maximums above the default threshold on the smoothed signal. With the same detection threshold, ChIPSeqR detects by default 57 725 nucleosome binding sites and 2 206 180 if asking for sub-peak detection, very far from the expected magnitude of 10^4-10^5 calls, according to yeast genome size and putative nucleosome length. A visual comparison of the mentioned methods is available as Supplementary Material (see Supplementary Fig. S1). Again, a benchmark analysis showed that nucleR was able to recover more nucleosome positions in synthetic data (95% for nucleR versus 81% for ChIPseqR) and to reproduce with higher definition the experimental coverage map ($P=0.29$ for nucleR versus $P=0.03$ for ChIPSeqR) (see Supplementary Material for details).

nucleR has a computational complexity between $O(N)$ and $O(N \log N)$. The package is accessible free of cost under LGPL-3 license scheme from our website (http://mmb.pcb.ub.es/nucleR) and the Instituto Nacional de Bioinformática site (http://www.inab.org) It should also be available on Bioconductor upon publication.

ACKNOWLEDGEMENTS

We thank Carles Fenollosa for testing the software and Özgen Deniz for the experimental support.

Funding: Spanish Ministry of Science and Innovation (BIO2009-10964 and Consolider E-Science); Instituto de Salud Carlos III (INB-Genoma España and COMBIOMED RETICS); Fundación Marcelino Botín.

Conflict of Interest: none declared.

REFERENCES

Chen,X. *et al*. (2010) A dynamic Bayesian network for identifying protein-binding footprints from single molecule-based sequencing data. *Bioinformatics*, **26**, i334–i342.

Deniz,Ö. *et al*. (2011) Physical properties of naked DNA signal gene regulatory regions. (in press).

Di Gesù,V. *et al*. (2009) A multi-layer method to study genome-scale positions of nucleosomes. *Genomics*, **93**, 140–145.

Jiang,C. and Pugh,B.F. (2009) Nucleosome positioning and gene regulation: advances through genomics. *Nature Rev. Genet.*, **10**, 161–172.

Lee,W. *et al*. (2007) A high-resolution atlas of nucleosome occupancy in yeast. *Nature Genet.*, **39**, 1235–1244.

Smith,S.W. (1999) *The Scientist and Engineer's Guide to Digital Signal Processing*, 2nd edn. California Technical Publishing, San Diego, CA, USA.

Yassour,M. *et al*. (2008) Nucleosome positioning from tiling microarray data. *Bioinformatics*, **24**, i139–i146.

Yuan,G.-C. *et al*. (2005) Genome-scale identification of nucleosome positions in S.cerevisiae. *Science*, **309**, 626–630.

3.2. htSeqTools: high-throughput sequencing quality control, processing and visualization in R

This library is a suite of different algorithms to perform quality control and general downstream analysis of high-throughput sequencing experiments. I used it for the first time when it was a library for internal use in the IRB Biostatistics core facility, during a lab rotation on my first year of the PhD. It allows the detection of inefficient ChIP-Seq experiments, the assessment of coverage uniformity or correction of overrepresented short reads due to a differentially effective PCR amplification (Fig. 34) among others. All the implemented methods feature strong statistics background and can cover different techniques based on sequencing. My collaboration with the final version of this library was in the debugging and optimization of the performance of some functions.

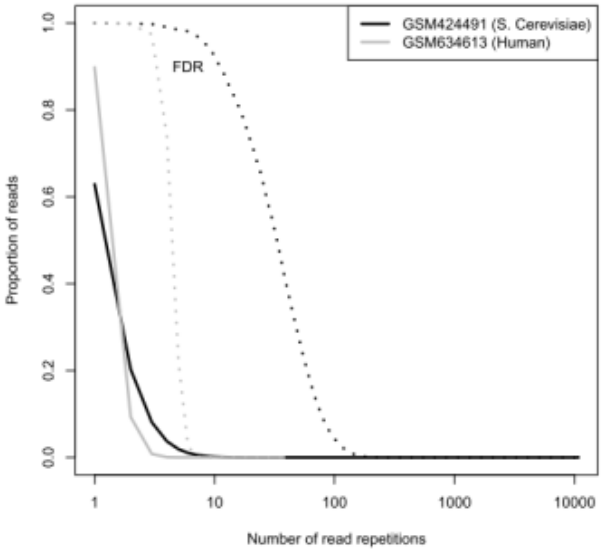


Fig. 34. Detection of over-amplification artifacts. The proportion of reads is plotted relative the number of reads repetition, and we observe that most of the reads are repeated less than 10 times in both organisms, but a very small portion of them can be repeated more than 10000 times. With this method, reads repeated more than the threshold marked by the False Discovery Rate (FDR) will be removed.

Publication:

Planet, E., Stephan-Otto, C., Reina, O., Flores, O. and Rossell, D. (2012) "*htSeqTools: high-throughput sequencing quality control, processing and visualization in R*", Bioinformatics (Oxford, England) 28: 589–90.

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

BIOINFORMATICS APPLICATIONS NOTE

Vol. 28 no. 4 2012, pages 589–590
doi:10.1093/bioinformatics/btr700

Sequence analysis

Advance Access publication December 22, 2011

htSeqTools: high-throughput sequencing quality control, processing and visualization in R

Evarist Planet[1], Camille Stephan-Otto Attolini[1], Oscar Reina[1], Oscar Flores[2] and David Rossell[1,*]

[1]Biostatistics and Bioinformatics Unit, Institute for Research in Biomedicine of Barcelona, Barcelona and [2]IRB-BSC Joint Research Program on Computational Biology, IRB Barcelona, Barcelona, Spain

Associate Editor: John Quackenbush

ABSTRACT

Summary: We provide a Bioconductor package with quality assessment, processing and visualization tools for high-throughput sequencing data, with emphasis in ChIP-seq and RNA-seq studies. It includes detection of outliers and biases, inefficient immunoprecipitation and overamplification artifacts, *de novo* identification of read-rich genomic regions and visualization of the location and coverage of genomic region lists.

Availability: www.bioconductor.org

Contact: david.rossell@irbbarcelona.org

Supplementary information: Supplementary data available at *Bioinformatics* online.

Received on July 13, 2011; revised on December 15, 2011; accepted on December 19, 2011

While analysis strategies for high-throughput sequencing data are proliferating, there remains a need for quality assessment, data processing and visualization methods. We provide tools to detect the presence of outliers, inefficient immuno-precipitation (IP), over-amplification and strand-specific biases. We implement strategies to adjust for these biases. Also, we provide routines for quick data formatting, analysis and visualization. `htSeqTools` is integrated in Bioconductor (Gentleman *et al.*, 2004), an environment offering a wide variety of analysis strategies. We take advantage of parallel computation and operations implemented in other packages to deliver computationally efficient solutions. `htSeqTools` can be a valuable complement for pipelines and advanced analysis strategies.

Below we show the main software features and use a *Saccharomyces cerevisiae* (GSE16926) and a human (GSE25836) ChIP-seq experiment as examples (www.ncbi.nlm.nih.gov/geo). The Supplementary Material contains a more detailed description, comparisons with existing approaches, typical workflows, as well as 2 ChIP-seq and 1 RNA-seq additional examples.

1 QUALITY CONTROL

• Visualize sample correlations: principal component analysis (PCA) is useful to assess quality and identify problematic samples. Unfortunately, it is not directly applicable to sequencing data. Instead, we measure the distance in read coverage between samples i and j as $d_{ij} = 0.5(1-\rho_{ij})$ where ρ_{ij} is the Pearson, Spearman or Kendall correlation between their log(coverage + 1). The log-scale reduces the influence of extremely high-coverage regions. We display d_{ij} in 2–3 dimensions via multi-dimensional scaling (MDS), so that Euclidean distances between points approximate d_{ij}. Figure 1 shows an MDS plot for the ChIP-seq experiment GSE25836. The distance between FOXA2 IP samples and their inputs is larger than the distance between replicates, indicating a satisfactory quality.

• Remove overamplification artifacts: PCR overamplification causes some reads to repeat an abnormally large number of times, which can induce biases in downstream analyses. Simultaneously, naturally occurring read repeats are expected. For instance, short genomes or IP samples typically show more read repeats than longer genomes or control samples, as they focus on smaller genomic regions. We model the number of repeats as a mixture of truncated negative binomial distributions [number of components set to minimize the BIC, Schwarz (1978)], and use an empirical Bayes approach akin to Efron *et al.* (2001) to estimate the False Discovery Rate (FDR). We fit the model after truncating 0.001 of the reads (by default) with highest number of repeats, as these are more likely to be artifacts. Figure 2 shows more read repeats in the *S.cerevisiae* than in the human data (solid lines). Reads with more than six repeats were flagged as overamplification artifacts in the human data at a 0.01 FDR, while for *S.cerevisiae* the cutoff was 138 repeats. The procedure adapts the cutoff to the nature of the data.

• Assess enrichment efficiency: in sequencing experiments such as ChIP-seq, MeDIP or DNase-seq, certain samples accumulate more reads in specific regions than their controls. A lack of such coverage variability can indicate sample preparation problems (e.g. inefficient IP) or a lack of pronounced peaks. We measure coverage inequality with the standard deviation (SD) and Gini's coefficient G (Gini, 1912), a classical econometrics measure of wealth inequality. The expected value of the coverage SD is proportional to $\sqrt{n}$ (see Supplementary Material), where n is the number of reads. The expected $E(G|n)$ also depends on n, but no closed-form expression is available. We estimate $E(G|n)$ by generating n reads uniformly distributed along the genome. In order to make samples with different n comparable, we report $SD_n = SD/\sqrt{n}$ and $G_n = G - E(G|n)$. Table 1 shows higher SD_n and G_n in the IP samples than in their respective controls, suggesting no sample preparation problems. Samples sequenced with GAII present clearer peaks than GAI samples, thus indicating an improvement in the technology.

*To whom correspondence should be addressed.

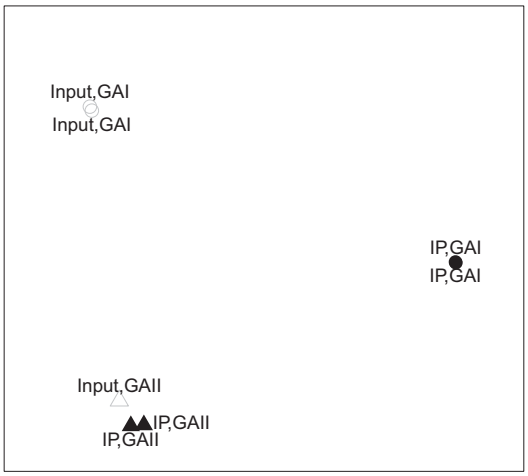


Fig. 1. GSE25836 MDS approximating log-coverage Pearson distances.

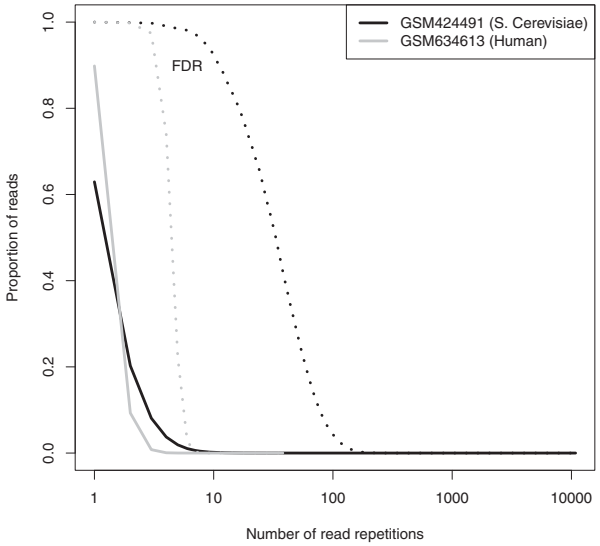


Fig. 2. Removing overamplifications. Dotted lines: estimated proportion of non-overamplified reads (FDR).

Table 1. Coverage SD_n and G_n for dataset GSE25836

ID	SD_n	G_n	Antibody	GA
GSM634613	0.115	0.0060	FOXA	GAI
GSM634615	0.114	0.0059	FOXA	GAI
GSM634617	0.110	0.0017	Input	GAI
GSM634619	0.106	0.0017	Input	GAI
GSM634614	0.124	0.0142	FOXA	GAII
GSM634616	0.135	0.0211	FOXA	GAII
GSM634618	0.120	0.0022	Input	GAII

• Correct strand bias: ChIP-seq fragment sizes cause reads on the $\pm$ strands to be shifted with respect to each other. With single-end reads this poses a challenge, as the fragment size distribution is unknown. Akin to Zhang *et al.* (2008), we scan for reads in high coverage regions and estimate the shift $\hat{s}$ as the mean distance between reads in the $+$ and $-$ strands. We add/subtract $0.5\hat{s}$ to the location of reads on the $\pm$ strand, respectively.

2 ANALYSIS

• Find read-rich regions: in many sequencing experiments, the goal is to identify *de novo* genomic regions of interest, e.g. binding sites, previously unannotated short RNAs or copy number variations. Although many analysis strategies are available, the computational burden of applying them to the whole genome is often excessive. It is therefore convenient to prescreen and focus the analysis. We implement a screening tool to detect all genomic regions with coverage above a user-specified threshold, count the number of reads in each region, and optionally compare the number of reads across samples via likelihood ratio or permutation chi-square tests. We also allow for refined peak calling within the selected regions.

• Visualize hits: we facilitate the visualization of a list of genomic regions by plotting the distribution of their distances to the closest gene/feature (in base pair or relative to the feature length) and average coverage profiles. Often it is useful to scan the genome for regions accumulating a large number of hits, e.g. peaks in ChIP-seq or differential expression in RNA-seq may reveal common regulatory mechanisms. We provide functions to detect and plot such areas.

Funding: Instituto de Salud Carlos III (INB-Genome España and COMBIOMED RETICS) funds (to O.F.); IRB Barcelona funds (to E.P., C. S.-O. A., O. R. and D.R.).

Conflict of interest: none declared.

REFERENCES

Efron,B. *et al.* (2001) Empirical Bayes analysis of a microarray experiment. *JASA*, **96**, 1151–1160.

Gentleman,R. *et al.* (2004) Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Gini,C. (1912) *Variabilita e Mutabilita*. C. Cuppini, Bologna.

Schwarz,G.E. (1978) Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464.

Zhang,Y. *et al*. (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R173.

3.3. DASiR: Programmatic data retrieval from DAS servers in R

Distributed Annotation System (DAS) is a web service protocol based on simple technologies. Databases available on the servers are queried with a simple HTTP request and the output is presented as an XML document. Thanks to its simple interface, DAS interfaces are broadly present in different servers and tools, as genome browsers. A complete list of more than 1500 DAS servers can be found in the URL http://www.dasregistry.org

Despite DAS servers are a valuable source of biological information, R/Bioconductor framework did not have any convenient interface to them, forcing the user to manually build the different queries and parsing the resulting XML.

With this convenient library, we wanted to facilitate the integration of information in R extending the connection capabilities of this framework. This is especially useful in the case of the University of California-Santa Clara (UCSC) Genome Browser, main data source from different genomic projects, like ENCODE or Roadmap Epigenomics, and which only provides a DAS interface for programmatic data access.

Despite this library was peer-reviewed before its publication in the Bioconductor repository, it has not been submitted to a scientific journal due to its technical scope. Anyway its broad applicability is proven by more than 2600 downloads within its first year from its publication (Apr 2013 – Apr 2014). Here, I present the *vignette* (manuscript demonstrating different use cases) available in Bioconductor.

Publication:

Flores, O. and Mantsoki, A. (2013) "DASiR: Programmatic data retrieval from DAS servers in R," Bioconductor (on-line)

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

Programmatic retrieval of information from DAS servers using the **DASiR** package

Oscar Flores Guri and Anna Mantsoki
Institute for Research in Biomedicine & Barcelona Supercomputing Center
Joint Program on Computational Biology

June 14, 2013

1 Introduction

Distributed Annotation System (DAS) is a protocol for information exchange between a server and a client. Is widely used in bioinformatics and the most important repositories include a DAS server parallel to their main front-end. Few examples are "UCSC", "Ensembl" or "UniProt".

The DASiR package provides a convenient R-DAS interface to programmatically access DAS servers available from your network. It supports the main features of DAS 1.6 protocol providing a convenient interface to R users to a huge amount of biological information. DAS uses XML and HTTP protocols, so the server deployment and client requirements are significantly less than MySQL or BioMart based alternatives. You can find an browsable list with more than 1500 on-line DAS servers in the url `http://www.dasregistry.org`.

Despite the DAS protocol supports querying different kinds of data, DASiR has been designed with ranges-features in mind. Querying genomic sequences and protein structures is also supported, but you probably will find better ways to access such data in R if you requiere an intensive use of them (Biostrings genomes or Bio3dD package for PDB structure, for example).

Here you have a brief summary of the main functions of DASiR:

- Set/get the DAS server: `setDasServer`, `getDasServer`
- Retrieve the data sources: `getDasSource`, `getDasDsn`
- Retrieve the entry points and types: `getDasEntries`, `getDatTypes`
- Retrieve information about the features: `getDasFeature`
- Nucleotide/amino acid sequence retrieval: `getDasSequence`
- Protein 3D structure retrieval: `getDasStructure`

For detailed information about these functions and how to use them refer to the DASiR manual or give a look to the following sections for an overview.

2 DAS metadata information handling

As an important part of the dialog between the client and the server consists on knowing which information has every server, DASiR provides the following functions to query metadata from a DAS server.

The first step before start quering and retrieving information is, of course, set the DAS server we will use during this session. Notice that DASiR only supports one active DAS session per R instance.

To set the server we will use the function `setDasServer`:

```
> setDasServer(server="http://genome.ucsc.edu/cgi-bin/das")
> getDasServer()
```

```
[1] "http://genome.ucsc.edu/cgi-bin/das"
```

The function `getDasSource()` will return the id's, the titles and the capabilities of the data sources available in the server in the form of a data frame. This function is important since it is necessary to use the exact name ("id" or "title" depending on the server) as reference name in the other functions of the package.

```
> #sources = getDasSource() #This will fail for UCSC (but no for ENSEMBL)
> sources = getDasDsn() #This will fail for ENSEMBL (but no for UCSC)
> head(sources)
```

```
[1] "hg19"    "hg18"    "hg17"    "hg16"    "panTro4" "panTro3"
```

We should also take into account that the values in capabilities of each data source returned by this function. If we query an unknown name or we query for a capability that is not implemented in the server (for example we ask for a atomic structure to a sequence server) DASiR will return a NULL value, but due to the nature of HTTP based queries, we cannot detect the origin of this error without overloading the server with cross-calls.

Despite "sources" is the common way to recover server features, some servers (as UCSC) are still using the deprecated "dsn" option to get the ids of the different datasets. The rule of a thumb is that if the first don't work, try the second.

Once we have the name of the database we want to query, we need to know where we can look. This is called "entry point" and could be from a protein ID to a chromosome number (depending on the type of server, of course). Notice that output will be a GRanges by default (if the server supplies start, stop and id values). If this is not possible or textttas.GRanges=FALSE, the output will be a data.frame.

```
> source = "sacCer3"
> entries = getDasEntries(source, as.GRanges=TRUE)
> head(entries)
```

```
GRanges with 6 ranges and 2 metadata columns:
      seqnames       ranges strand | orientation subparts
         <Rle>    <IRanges>  <Rle> |    <factor> <factor>
  [1]       IV [1, 1531933]      * |           +       no
  [2]       XV [1, 1091291]      * |           +       no
  [3]      VII [1, 1090940]      * |           +       no
  [4]      XII [1, 1078177]      * |           +       no
  [5]      XVI [1,  948066]      * |           +       no
  [6]     XIII [1,  924431]      * |           +       no
  ---
  seqlengths:
      I   II  III   IV   IX    M    V   VI ...    X   XI  XII XIII  XIV   XV  XVI
     NA   NA   NA   NA   NA   NA   NA   NA ...   NA   NA   NA   NA   NA   NA   NA
```

Finally, the different attributes we can ask for, called "types" could be obtained with the function `getDasTypes`.

```
> types = getDasTypes(source)
> head(types)
```

```
[1] "blastHg18KG"      "est"              "intronEst"        "mrna"
[5] "ensGene"          "esRegGeneToMotif"
```

Think about Sources as the name of the database, Entries are the tables of this database and finally Types are the columns on this table.

In general a `character` vector is returned for most of the functions, except for `getDasSource` and `getDasEntries` which return a data.frame with the name of the entry the ranges of the elements and possibly other attributes depending on the server.

3 Querying features

The `getDasFeature` function queries the DAS server and returns the available information for the type(s) at the given range(s).

```
> ranges=entries[c(1,2)]
> types=c("sgdGene","mrna")
> features = getDasFeature(source, ranges, types)
> head(features)
```

```
  segment.range                     id    label type method  start    end score
1             1 AY169693.chrIV.84188.0 AY169693 mrna   BLAT  84189  85450   990
2             1 AY169693.chrIV.84188.1 AY169693 mrna   BLAT 520477 520627   990
3             1 AY169693.chrIV.84188.2 AY169693 mrna   BLAT 520628 520633   990
4             1 AY169693.chrIV.84188.3 AY169693 mrna   BLAT 520634 520643   990
5             1 AY169693.chrIV.84188.4 AY169693 mrna   BLAT 540324 540328   990
```

```
6             1 AY169693.chrIV.84188.5 AY169693 mrna   BLAT 593167 593177   990
  orientation phase                        group
1           +     - Link to UCSC BrowserAY169693
2           +     - Link to UCSC BrowserAY169693
3           +     - Link to UCSC BrowserAY169693
4           +     - Link to UCSC BrowserAY169693
5           +     - Link to UCSC BrowserAY169693
6           +     - Link to UCSC BrowserAY169693
```

A `data.frame` is returned with the contents of the specific annotation for the given ranges and types in the server (again, note that servers could provide different/additional information for types with the same name).

4 Querying nucleotide or amino acid sequence

The `getDasSequence` function queries the DAS server and retrieves the nucleotide or amino acid sequence for the given ranges.

```
> #Now we will retrieve sequences from VEGA server
> setDasServer("http://vega.sanger.ac.uk/das")
> source = "Homo_sapiens.VEGA51.reference"
> ranges = GRanges(c("1","2"), IRanges(start=10e6, width=1000))
> #Returning character vector, we only ask 50 first bases in the range
> sequences = getDasSequence(source, resize(ranges, fix="start", 50))
> print(sequences)
```

```
[1] "aaccccgtctctacaataaattaaaatattagctgggcatggtggtgtgt"
[2] "gtattagtttgtttccacactactatgaagatactacctgagactgggta"
```

An `character` vector is returned (the default class) containing the nucleotide or amino acid sequences that match the content of the annotation for the given ranges in the server. Automatic conversion to Biostring classes is supported with the `class` attribute:

```
> #Now we specify we want a AAStringSet (Biostring class for AminoAcids strings)
> #and query for the whole sequence length
> sequences = getDasSequence(source, ranges, class="AAStringSet")
> print(sequences)
```

```
  A AAStringSet instance of length 2
    width seq
[1]  1000 aaccccgtctctacaataaattaaaatattagct...aattttccagtaaaagtcaaggcaaacccaaga
[2]  1000 gtattagtttgtttccacactactatgaagatac...tagactgagaggagagaaattggagacaacaaa
```

5 Query a protein 3D structure

The `getDasStructure` function queries the DAS server and retrieves the 3D structure, including metadata and coordinates for the given query (ID of the reference structure) using the data source id (or title).

```
> #On 2013-03-05 there are 2 structure servers in dasregistry.org...
> setDasServer(server="http://das.sanger.ac.uk/das")
> #Get the sources with "structure" capability...
> sources = getDasSource()
> sources[grep("structure", sources$capabilities),]
```

```
        id     title   capabilities
2 structure structure das1:structure
```

```
> source="structure" #...which name is also "structure"
> query="1HCK" #PDB code
> structure=getDasStructure(source,query)
> head(structure)
```

```
  atomID atomName       x       y      z  type groupID name
1      1      N   102.329 111.862 92.452 amino       1  MET
2      2     CA   103.332 112.165 93.516 amino       1  MET
3      3      C   103.877 113.584 93.255 amino       1  MET
4      4      O   103.802 114.075 92.129 amino       1  MET
5      5     CB   104.437 111.099 93.495 amino       1  MET
6      6     CG   105.176 110.881 94.812 amino       1  MET
```

An `data.frame` is returned with the information of every atom of the reference structure of the given query. Notice that a single residue (identified by groupID and name) has few atoms inside. Depending on your purposes, you can find this way to represent the structural information unconvenient (for example, if you are used to work with PDB format). Structure query in DAS seems a secondary feature of the protocol (as only 2 servers out of >1500 are supporting it) but anyway we wanted to support the maximum amount of features that DAS protocol implement. We realize that are better options to work with PDB files, but maybe this one can help you for quick structural analysis.

6 Plotting of obtained features

The `plotFeatures` function creates a basic plot with the features retrieved by `getDasFeature` or adds them to an existing plot.

```
> #Let's retrieve some genes from UCSC Genome Browser DAS Server now
> setDasServer(server="http://genome.ucsc.edu/cgi-bin/das")
> #Official yeast genes and other annotated features in the range I:22k-30k
```

```
> source = "sacCer3" #Saccharomices Cerevisiae
> range = GRanges(c("I"), IRanges(start=22000, end=30000))
> type  = c("sgdGene", "sgdOther") #This is also the name of the UCSC tracks
> features = getDasFeature(source, range, type)
> #Only the main columns
> head(features[,c("id", "label", "type", "start", "end")])
```

```
                      id      label     type start   end
1 YAL063C-A.chrI.22394.0  YAL063C-A  sgdGene 22395 22685
2   YAL063C.chrI.23999.0    YAL063C  sgdGene 24000 27968
3  YALWdelta1.chrI.22230 YALWdelta1 sgdOther 22231 22552
4        FLO9.chrI.24000       FLO9 sgdOther 24001 27968
```

```
> plotFeatures(features, box.height=10, box.sep=15, pos.label="top", xlim=c(22000,30000))
```

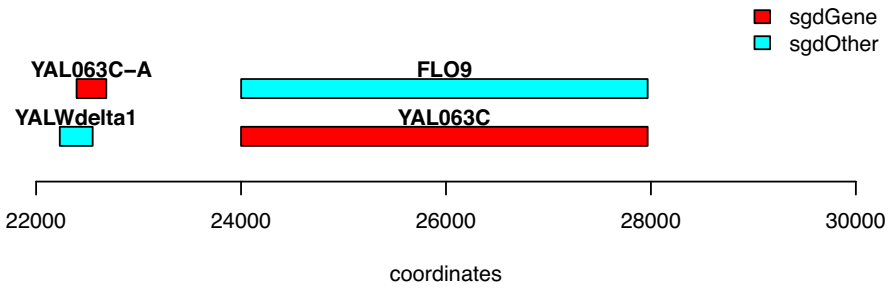


Figure 1: Basic feature plot generated with `getDasFeature` and `plotFeatures`

The `plotFeatures` function is a simple way of drawing text boxes in a new or an existing plot with the features retrieved. Notice that when overplotting to a already open graphical device, the x-coordinates must match.

And this is all. You can find more detailed information and examples of each function in the R manpages for DASiR. We hope this package helps you in your data mining.

3.4. Dynamic analysis of nucleosome positioning at read level

Although nucleR (page 109) enables a fast, intuitive and discreet analysis of the nucleosome positioning, one of the biggest challenges in the analysis of the nucleosome maps is the comparison between two given experiments. In the different articles presented in this thesis we found this need, which we addressed by using some type of correlation, standard deviation or coverage comparison. Despite these metrics provide a measure of the magnitude of change in a given region, they don't explain the sense of the changes (Fig. 35 a, b c).

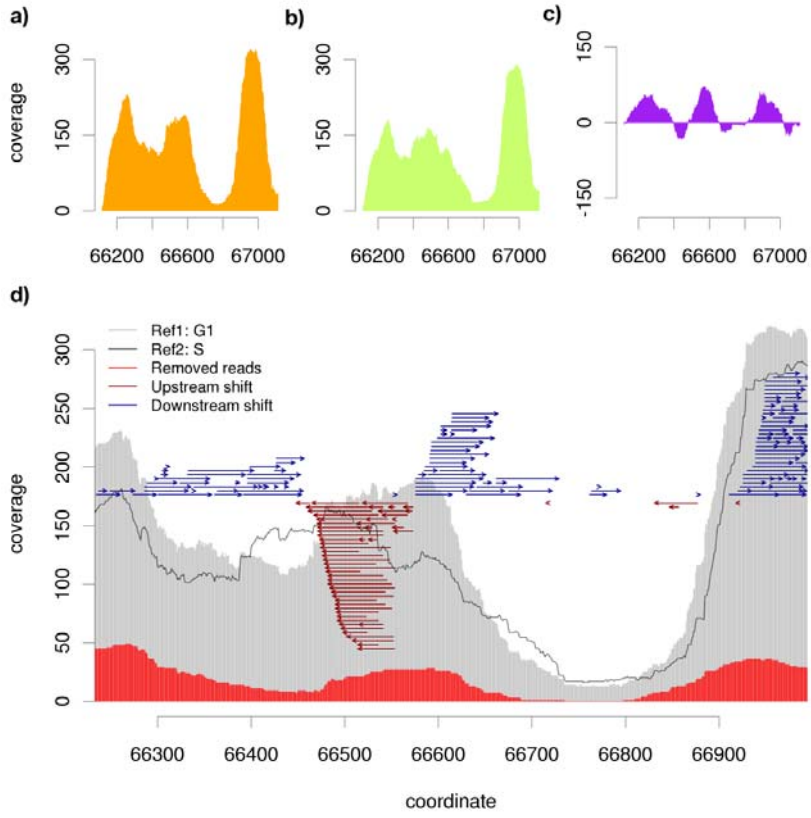


Fig. 35 a,b) Coverage profiles from two different nucleosome maps of a given region in the yeast genome. c) The base-by-base difference of them provides a quantification of the most changing regions, but does not explain their cause. d) Read-level annotation of Nucleosome Dynamics provide a low-level annotation of the changes, which show that the nucleosomes around the position 66600 are dispersed both upstream and downstream which explains the change in the coverage profile

The idea of Nucleosome Dynamics (name inspired by Molecular Dynamics simulations) was to calculate the changes that individual nucleosome molecules (short-reads) need in one reference experiment in order to reproduce the coverage profile observed in another reference experiment. The aggregation of these changes will allow the inference of global chromatin changes providing both a magnitude and sense. Basic read-level operations could be shifts, insertions or deletions and changes in the size of the reads (over/under digestion). At the same time, combining some of those basic operations we can give place to combined dynamics as dispersions, concentrations, opening or closing of nucleosome architectures (Fig. 35 d).

Regarding the methodological part, the calculation of the changes at the read level was implemented as an expectation-maximization problem solved by an *ad-hoc* implementation of a genetic algorithm. Despite this library is implemented in R —taking advantage of all the sequencing-related libraries in the Bioconductor framework— the core of the program was implemented in low-level C code to reduce the overhead in the most critical parts of the algorithm. Additionally, a relevant work has been done regarding the algorithmic parts. For example, we needed a function in order to detect cross-shifting of two reads (causing redundant moves). A naïve implementation of this problem supposes a complexity $O(n^2)$, being *n* the number of reads taken into account. With an smart implementation, we can obtain a similar measure by separating the reads by the sense of the shift (upstream or downstream) and calculating the difference in the "coverage" of the modulus of the shift. This is an operation of cost O(n+w), being *w* the size of the window accounted in the simulation. Taken together, the algorithmic optimizations of this library provided a gain in the execution time of 200X compared to the initial straightforward implementation (from 6h to 5 minutes in a single-core execution with default parameters).

In the following article, Nucleosome Dynamics we present its use and illustrate it with 3 use cases: a tour-de force calculating hotspots in a full human genome (Fig. 36), an analysis of the changes in chromatin featured in the nucleosome maps of another article featured in this thesis, *Hog1 bypasses stress-mediated down-regulation of transcription by RNA polymerase II redistribution and chromatin remodeling*, and the animation of the nucleosome changes in the promoter of a gene related to cell cycle.

Despite the fact that both the program and the manuscript were virtually finished, the appearance of a similar method for the detection of nucleosome dynamics in the journal Genome Research

compromised the novelty of our method (Chen *et al.*, 2013). Therefore, we decided to better explode our powerful read-level analysis to include new dynamics operators and improve the algorithmic, still aiming to a high impact journal. The public version of the manuscript below and the Nucleosome Dynamics code should be ready in the subsequent months to this thesis.

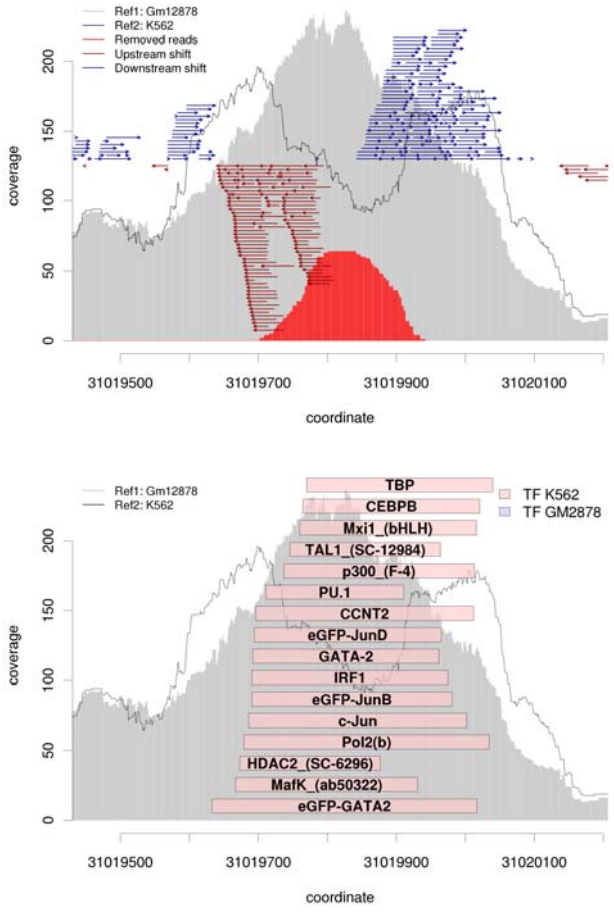


Fig. 36. Nucleosome Dynamics reveal a large eviction of nucleosomes in the center of the image (top) due to the binding of different experimentally annotated transcription factors (bottom). The reference nucleosome maps used here are from GM12878 and K562 cell lines, both obtained from ENCODE (Dunham *et al.*, 2012)**.**

Publication:

Flores, O. and Orozco, M. "Dynamic analysis of nucleosome positioning at read level" (in preparation)

Dynamic analysis of nucleosome positioning at read level

Oscar Flores[1,2] and Modesto Orozco [1,2,3*]

[1] Institute for Research in Biomedicine and Barcelona Supercomputing Center. Joint Research Program on Computational Biology. Baldiri Reixac 10-12. 08028 Barcelona, Spain;

[2] National Institute of Bioinformatics, Structural Bioinformatics Node, Baldiri Reixac 10-12, 08028 Barcelona, Spain

[3] Department of Biochemistry and Molecular Biology. University of Barcelona, 08028 Barcelona, Spain;

* Corresponding author: Prof. Modesto Orozco, modesto.orozco@irbbarcelona.org

Running head: Dynamic analysis of nucleosome positions at read level

Keywords: nucleosome positioning, chromatin dynamics, automatic annotation

ABSTRACT

The study of nucleosome maps over the past decades has revealed that the regulatory role of chromatin in regulatory regions goes beyond the simple occlusion of DNA according transcriptional states. Recent works point to the need of studying nucleosome variability accounting for specific effects on cell populations rather than capturing only broad, general trends. Parallel to this, advances in the sequencing technology and computational power allows formulating new questions which were technically impossible till now. Here, we present Nucleosome Dynamics, an algorithm which enables the comparative analysis of nucleosome positioning at single read level, allowing fine characterization of observable changes providing also an explanation of their cause in a dynamic context. The read-level annotation enables the analysis of an unprecedented resolution in terms of nucleosome inclusion, eviction, shift or changes in the digestion of the chromatin. We exemplify its application in three different cases: the characterization of nucleosome differences between two human cell lines, the quantification of the changes suffered between wild-type and HOG1 mutant yeast cells under osmostress and the representation of changes during different stages of cell cycle in yeast. This software is implemented over the R/Bioconductor framework and is available in the web http://mmb.irbbarcelona.org/nucdyn/

INTRODUCTION

Nucleosomes are the basic organizational unit of the chromatin, allowing a high compaction of the DNA in the nuclei of the cell[1]. In a lineal representation of the chromatin, nucleosomes are usually thought as *beads-on-a-string*, where the nucleosomal DNA wrapping the histone core is separated by around 20bp of linker DNA. In the study of genome-wide nucleosome coverage, current experimental technologies provide us averaged representations of different cells populations. This causes that a single region can show different degrees of variation in the nucleosome positions, causing phased of fuzzy nucleosomes arrays [2]. In the last years, sequencing technologies provided nucleosome maps of increased resolution

and coverage at decreasing costs and today they are evolving towards the single-molecule scanning, providing data of unprecedented quality for single genomes [3].

Different methods for the unambiguous classification of genomic regions in function of their nucleosome occupancy (nucleosome callers) have been proposed[4–7] during the last years, including our fast and non-parametric method nucleR[8]. Despite the undeniable usefulness of those methods for a coarse-grained classification of the nucleosomal states[9,10], recently new methods for the analysis of subpopulations of nucleosome distributions have been proposed with the focus on understanding the dynamics of chromatin with an increased level of detail[11–13]. Also, new methods for a coverage-level annotation of changes have been proposed [14].

In this work we introduce Nucleosome Dynamics, an R/Bioconductor package for the analysis of chromatin changes between two reference maps at single read level. This is, in our knowledge, the first algorithm to use short-read level annotations in the comparative analysis of nucleosomes, allowing an unprecedented power for the understanding of chromatin in dynamic conditions. The differences of coverage between two nucleosome maps in a specific region can be explained by different experimental conditions between subpopulations of cells yielding to a reorganization of apparent nucleosome coverage peaks (Figure 1). Nucleosome Dynamics allows the detection of displacements, inclusions, evictions and changes in the read width caused by differential digestion of chromatin between two reference maps. This read-level annotation not only allows the description of the observable changes in the coverage profiles, but also provides information about the sense and magnitude of the changes, providing a framework for the high-resolution integrative analysis of chromatin variation.

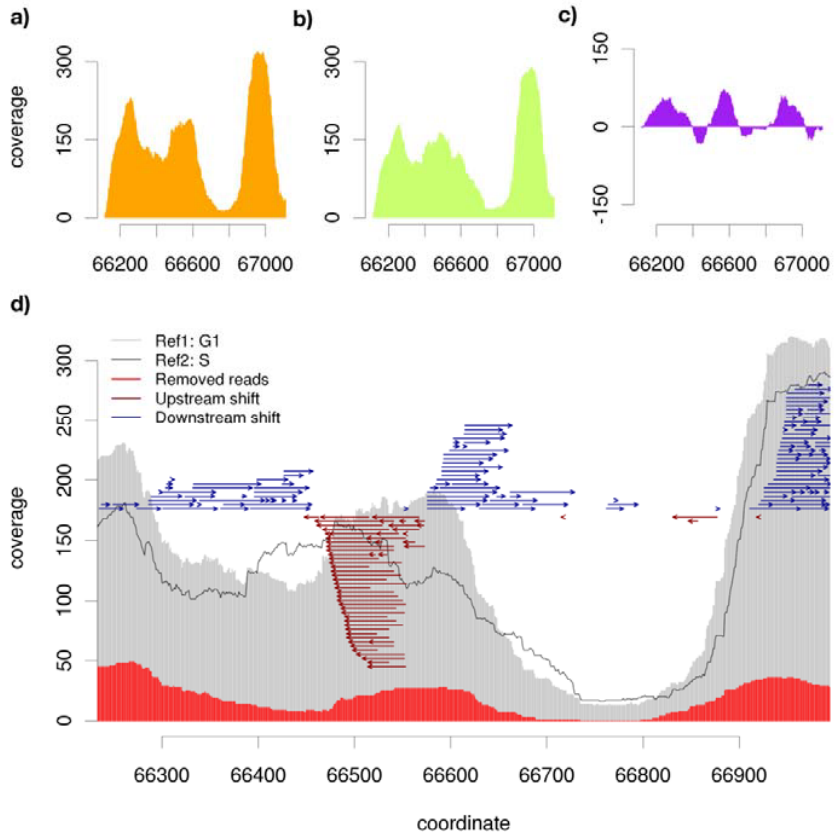


Figure 1. a) Coverage around -500:500 bases around YPL256C TSS locus for Reference map 1 (orange) **b)** for Reference map 2 (green) and **c)** the base-to-base difference for both references. The pink area is a measure of the change, but doesn't explain its causes**. d)** Dynamics profile of the same locus. Grey shaded area is the coverage profile of Ref1. Black dashed line is the coverage profile for Ref2. Reads shifting are shown as red (upstream) or blue (downstream). The changes in the coverage due to read eviction are shown as a red profile at the bottom of the figure. Dynamics view show that the coverage change around 66500-66600 coordinates are largely caused by a displacement of the reads in that region. Dataset used here[10] for Reference map 1 (a) corresponds to G1 stage of the Cell Cycle and map 2 (b) to S stage.

RESULTS

Workflow and data pipeline. An overview of the workflow in Nucleosome Dynamics is shown in Figure 2. The main inputs to Nucleosome Dynamics are the two reference datasets containing the sequencing reads to be compared. Every reference dataset consists in one or more indexed BAM files previously aligned to the genome using third-party tools. Multiple input files for each reference provides an interface for sequencing replicates, providing an increment of the fold-coverage in large genomes.

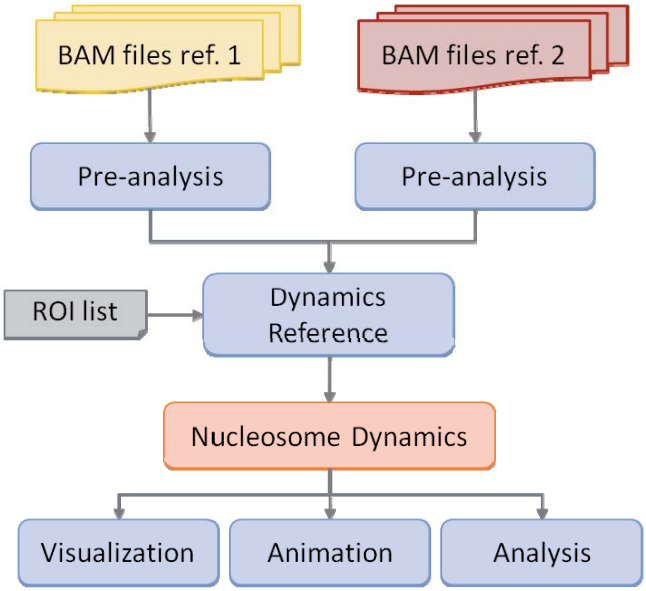


Figure 2. Workflow of Nucleosome Dynamics

For each different sample, we run a pre-analysis to obtain the metadata which will allow later normalization of the dataset and comparison between experiments with possibly technical differences. During the pre-analysis, different metrics as the fold-coverage or a statistical assessment of the PCR over amplification (see Methods) will be calculated. Those metrics avoid potential biases in comparisons between reasonably similar datasets (same organism, similar experimental conditions, minimum fold-coverage) independently of the specific technologies used in the sequencing (machine vendor, single or paired end sequencing, samples multiplexing, different fold-coverage, etc.).

Nucleosome Dynamics approach is based on an optimization problem (see next section). This makes the search space grown proportionally to the length of the regions studied. In order to maintain a compromise between the usability and the detail of the analysis, the algorithm is executed in user-defined windows with a recommended width between 500 and 1500 bases. Despite larger windows are possible, this provides a good compromise between analysis speed and resolution of the analysis in specific Regions of Interest (ROIs) such as binding sites, gene boundaries, regulatory regions, etc. Despite some users could be interested in running the dynamics in their own ROIs, we provide a function for the automatic genome-wide detection of variable regions based on a coarse-grained recognition of regions with differences in the coverage (see Methods).

The last step before the dynamics calculation itself is the generation of a standalone dynamics reference object for each of the ROI we will study. This object contains all the required information to execute independently each dynamic, allowing parallel processing of different regions in computer clusters without requiring access to the input files.

Nucleosome Dynamics. The calculation of the dynamics is based on an optimization problem solved by a genetic algorithm (see Methods). Starting from a set of reads Ref1, we want to find the best set of changes applied to every read that maximize its resemblance with Ref2. The read-level operations could be shifts (upstream or downstream displacements), deletions (nucleosome evictions), insertions (nucleosome inclusions) or changes in the width of the read (over-/underdigestion). The score of a given configuration is obtained by comparing the modified coverage of Ref1 with the coverage in Ref2. In order to constrain the results and keep the biological sense of the simulation, those solutions involving redundant or an unnecessary large amount of changes reads are penalized. The final score tries to keep a balance between the reads match and the number of changes applied. Despite this expectation-maximization approach precludes a single optimal solution, the analysis of the robustness for this method (see Robustness and performance) shows a high level of reproducibility and convergence between executions.

After the simulation, the resulting read-level annotation provides highly detailed data about the most probable source of variability. Although this enables the high-resolution analysis of the data, we also provide a method to reduce and summarize the outcome information, detecting hotspots in the dynamics. A hotspot is defined as a large accumulation of reads with common dynamics. Basic hotspots are *Shift*, *Insertion*, *Eviction, Overdigestion* and *Underdigestion* but loci sharing two or more hotspots can be combined in different ways (see Table 1 for a comprehensive list).

Basic hotspots	
Shift +	Reads shifting downstream
Shift –	Reads shifting upstream
Insertion	Reads added
Eviction	Reads removed
Overdigestion	Reads width decreased
Underdigestion	Reads width increased
Combinations	
Concentration	**Shift +** followed by **Shift –**
Dispersion	**Shift –** followed by **Shift +**
Opening –	**Eviction** overlapped with **Shift –**
Opening +	**Eviction** overlapped with **Shift +**
Opening <>	**Eviction** overlapped with **Dispersion**
Closing –	**Insertion** overlapped with **Shift –**
Closing +	**Insertion** overlapped with **Shift +**
Closing <>	**Insertion** overlapped with **Concentration**

Table 1. Basic and combined hotspots

For every hotspot, three measures are provided: 1) the absolute number of reads involved in the hotspot, 2) the relative amount of reads involved in the hotspot relative to the total number of reads in the dynamics window and 3) the relative amount of reads involved in the hotspot relative to the number of reads overlapping this position. The first two provide a measure of the magnitude of the

hotspot; meanwhile the last one allows a quantification of the relative number of cells which are affected by a change. For example, for a window containing 1000 reads a shift hotspot affecting 20 of them (absolute global number) will be reported as 2% respect the total number of reads in the window (relative global number). Additionally, supposing that in this specific position we find a total coverage of 40 reads, this will mean that 50% of reads in locus are affected (relative local number). This provides a quantifiable measure of the effect of in vivo factors, such as transcription factor binding or chromatin remodelers in the nucleosome population.

Description of nucleosome dynamics between GM12878 and K562 cell lines. Recently, the ENCODE consortium released one of the most comprehensive catalog of genomic elements in different human cell lines[15]. Nucleosome maps for the GM12878 (lymphoblastoid) and K562 (chronic myelogenous leukemia) cell lines were published as part of this dataset, as well as other associated data including transcription factors and epigenetic marks. The aggregated size of those nucleosome maps is 172Gbytes in two sets of nine BAM files providing a fold-coverage of 20-22 times the human genome. Here, we used this dataset as a *tour de force*, for Nucleosome Dynamics.

After discarding ambiguous regions (see Methods), we scanned a total of 1.7Gbases looking for regions with nucleosome coverage correlations lower than 0.3 in tiled windows of 1000bp with a 500bp spanning. Despite correlation-metric is highly sensitive to small fluctuations in the coverage and most users would prefer a global metric as regions with high coverage differences per base, we wanted to exemplify here a worse-case scenario. The computational cost of this process was a total 115 CPU hours with 3GHz processors and a memory peak of 11GBytes. After merging the overlapping windows and selecting its central 1000bp, we obtained a list of 1.2 million ROIs comprising a total of 1.2Gbases in the whole genome. Using default Nucleosome Dynamics parameters, the average execution time of a region in this dataset was 5.7 minutes per 1000 bases window. This implies that the cost of running Nucleosome Dynamics in the whole human genome (not only in a given ROI list) with the nucleosome maps provided by

ENCODE consortium would be around 100.000-120.000 CPU hours. Of course, with the use of modern clusters and parallelization the real time required can be reduced several orders of magnitude. This benchmark shows that the applicability of this method is not compromised by the computational power available nowadays, enabling the analysis at read level of a whole human nucleosome map in one month with a modern mid-size cluster. Of course this is an extreme case, as most of the users would be interested in a limited list of ROI or the analysis.

As an use case, we ran the dynamics in 8295 ROI detected in chromosome 21 with the and selected those with relevant hotspots (involving at least 20 reads with a relative proportion of 5% respect the total number of read in the window). This led to a set of 6130 dynamic regions where we analyzed the relationship between the changes in the nucleosome occupancy and *in vivo* factors, including the chromatin states[16] and 114 different Transcription Factors (TFs) present in both cell lines. In 713 of the 6130 dynamic regions (11.6%) we found at least one experimental TF Binding Site (TFBS) annotated for the specific cell line. The majority of those dynamic regions where annotated as active regulatory regions (20% strong enhancers, 13% weak enhancers, 8% active promoters, 7% insulators and 6% weak promoters), whereas only a small proportion was classified as heterochromatin (16%).

Dynamics descriptors in TFBS show, in general, a relevant change in the nucleosome coverage when TFs are bounded in one cell line but not in the other. The detailed view of one arbitrary region is presented in Figure 3. In Figure 3a a large opening hotspot (nucleosome dispersion + eviction) is found in the central coordinates. This hotspot affects 171 reads and represents a 24% of the reads present in this window. The relative amount of reads affected in the hotspot center is 61%, meaning that more than the half of the cells population in this area was affected by this change in chromatin. After analyzing the presence of TFs we found that no TFBSs were reported in this locus for GM2878, but K562 shows the binding of large complex containing, among others, transcriptional elements as TBP and Pol2 (Figure 3b). Additionally, the chromatin state for GM2878 in this area was classified as “heterochromatin”, meanwhile for K562 is “strong enhancer”.

Apart from the detailed analysis, the automatic annotation of nucleosome dynamics enables the programmatically analysis of the changes in the chromatin in presence of TFBS. We studied the effect of Pol2 binding, usually associated to an ongoing transcription. In general, dynamic regions with bounded Pol2 shown an enrichment around 7% (p-value=9.44e-06) of nucleosome movement rather than other dynamic regions Hotspot characterization show different trends in the case that Pol2 is reported for GM2878 (n=44) or for K562 (n=40) (cases with Pol2 bounded in the two cell lines have been omitted). As GM2878 was considered as Ref1 and K562 as Ref2 in our study, this constrains the directionality of the changes. Meanwhile for the cases where Pol2 is bound to GM2878 the hotspots are mainly "Inclusions", whereas in the cases where Pol2 is bound to K562 there is a high prevalence of "evictions" and "openings" (shifts combined with evictions). Our intention here is not achieve relevant conclusions rather than illustrate Nucleosome Dynamics in a large-scale experiment.

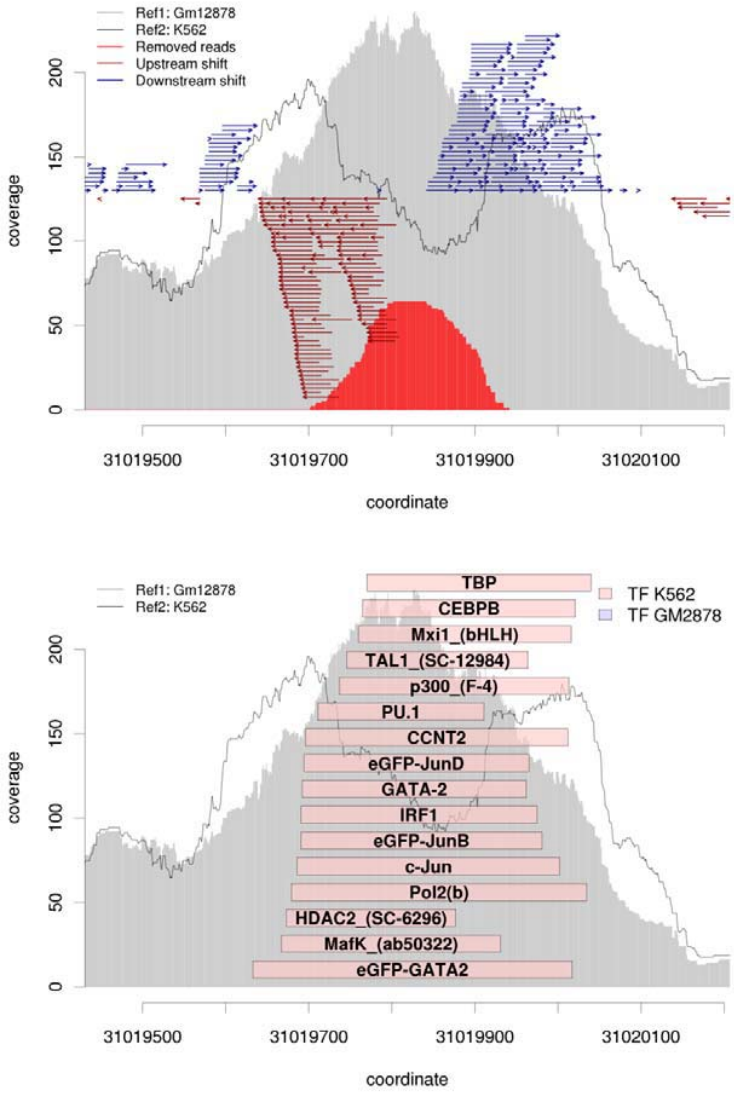


Figure 3. a) Dynamics of a region of the chromosome 21 with a massive opening in the center of the plot. **b)** Experimental TFBS annotated in this region overlapped with the coverage profiles of both samples.

Quantification of changes due to osmostress between wild type and hog1 mutant yeast strains. In a previous work[17], authors studied the Hog1 mediation in chromatin changes at stress-responsive loci. They compared the nucleosome coverage profiles of wild-type (WT) yeast (strain BY4741) and Hog1 mutant (YGM61) before and after osmostress (0.4M NaCl). In their work, the

authors found that Hog1-dependent genes show a dramatic change of the nucleosome occupancy after stress at both the promoter and ORF regions for the WT, meanwhile these changes were minimized in the Hog1 mutant. Despite the changes were evident in the combined plots of the coverage profiles around Transcription Start Sites (TSS) of Hog1-dependent genes, here we quantified and studied the dynamics in those loci.

We ran two sets of dynamics using the WT nucleosomal map and comparing it with the WT+NaCl map and with the Hog1Δ+NaCl later. The regions studied were the 1000 flanking bases (500 upstream and 500 downstream) around TSS of every gene annotated as Hog1-dependent in this work. We executed Nucleosome Dynamics with default parameters and performed an automatic search for hotspots. The results showed that the number of reads affected by some dynamic change was significantly larger (p-value = 10^{-9}) in the case WT vs WT+NaCl (average of 45% of the reads) than in the WT vs Hog1Δ+NaCl (average of 36% of the reads. Despite the numbers for shifts and inclusions hotspots are similar in the two cases, a relevant difference was presented in terms of eviction, with a greater prevalence in the case of the WT under osmostress than in the Hog1 mutant in the same conditions (Table 2). From these results we can numerically conclude that a significant change in the chromatin happens after osmostress between the two strains, which is mainly caused by a bigger eviction of nucleosomes, consistent with the differential role of chromatin remodelers stated in this work.

Hotspot	WT vs WT+NaCl	WT vs Hog1Δ+NaCl
SHIFT +	22,744 (03.34%)	21,992 (03.78%)
SHIFT -	20,560 (03.15%)	20,140 (03.66%)
EVICTION	310,978 (36.05%)	215,839 (25.72%)
INCLUSION	1,820 (01.20%)	1,604 (01.90%)

Table 2. Type and number of hotspost in the dynamics comparing WT and Hog1D. Numbers in cells represent the total number of reads affected, in brackets the relative fraction respect the number of reads considered.

Dynamic picture of the nucleosomal changes during the cell cycle in yeast. In the last example, we will use Nucleosome Dynamics for illustrating the changes in chromatin in a cell-cycle related gene. We studied in previous experiments performed in our group[10] the impact of yeast's cell cycle in the chromatin. In this example we applied Nucleosome Dynamics to the region flanking the TSS of the gene MFA1, which is related to α-factor pheromone mating and its desensitization response. In Figure **4** we illustrate the changes in the nucleosome coverage profiles between the G1 stage (0' after α-release) and late S stage (30' after α-release).

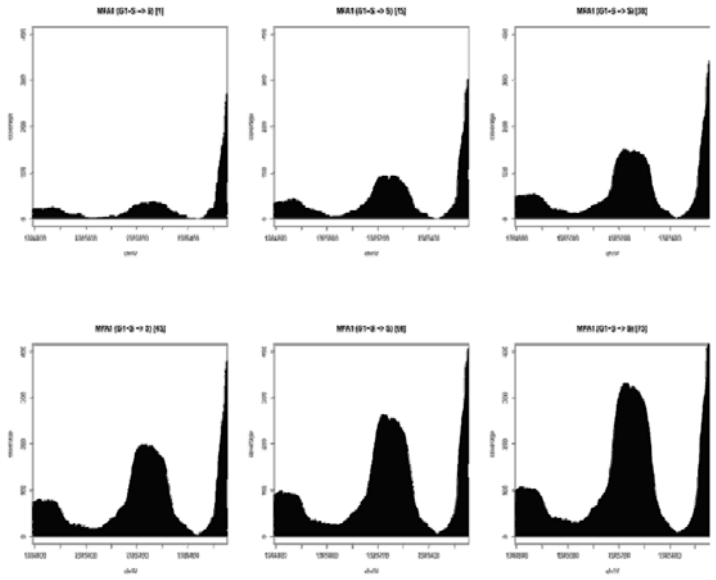


Figure 4. Different frames of the Nucleosome Dynamics animation calculated between the nucleosome profiles of the Transcription Start Site of the gene MFA1 between the G1 and S cell cycle stages.

The change in the chromatin profile shows a dramatic inclusion of the -1 and +1 nucleosomes around the Transcription Start Site, linked with a drastic drop of the expression as reported in this work. Apart from the analysis-tasks presented in previous examples, the dynamics interpolation provided in this software can generate custom animations of those changes for visualization purposes.

Robustness and reliability. Due to the stochastic nature of expectation-maximization algorithms the reproducibility of the results are not guaranteed by

construction. Therefore we studied the similarity of the generated dynamics in different executions of the algorithm with the same input parameters.

In order to evaluate the robustness of the dynamics predictions, we performed 1000 executions of 250 regions randomly sampled from the ENCODE dataset presented above. We detected the hotspots of those regions and analyzed their score in coverage and in shifts. Then we analyzed the different number of hotspots found in the same region. *(NOTE: preliminary results show that in more than 90% of the cases the same basic hotspots with a relative coverage greater than 1% are detected within a deviation of 10bp, but final results need to be calculated before the final publication.)*

Additionally to the robustness test we assessed the reliability of the dynamics predictions. One of the problems we found previously when trying to evaluate the performance of our nucleosome caller, nucleR[8], was the lack of a "gold standard" for validation. We solved this problem by running dynamics using synthetically generated nucleosome maps with perfect information of the different parameters describing the maps, as in other works[12,8].

We tested Nucleosome Dynamics in two different scenarios. In the first one we generated two synthetic nucleosome maps with 5 coverage peaks using the same random seed, so they share the same random distribution of reads. Then, as part of the generation method, we removed one of the putative peaks in the second map and run the dynamics using the read information that nucleR's synthetic maps generation provides. This process was repeated 1000 times changing the random seed in every execution, accounting for 1000 different nucleosome maps but all of them sharing a nucleosome eviction in some random position. In the 88.9% of the cases Nucleosome Dynamics was able to predict that eviction. In the cases where the detection was not successful we found that the "evicted" nucleosome had a coverage value lower than 5 reads. The coverage for each peak is, by default, a random value between 1 and 20, meaning that for a given putative peak position there can be up to 20 reads with slight variations in the position.

In the second case we generated two synthetic maps with 10 fixed putative nucleosome locations. In one of them, we shifted a random putative peak position between 20 and 50 bases upstream. We repeated this with the immediately

upstream putative nucleosome position adjacent to the previous one, but in this case the shift was applied in the downsteam direction, forcing a "concentration" pattern. In this case, the peak coverage was forced to be between 20 and 40 reads. Again, we executed Nucleosome Dynamics in 1000 random simulations with the described characteristics. In 86.4% of the cases the "concentration" hotspot was detected correctly. In the cases where detection was not successful, we found that the random shift applied in both peaks was too short, causing the magnitude of the combined "shift" being smaller than the threshold used by the hotspot detection to detect a "concentration" (a "shift +" followed by a "shift –"). In almost all cases the individual shifts were correctly annotated.

Computational performance. Despite most of the development time has been devoted to make this code as fast as possible, we cannot skip the fact that expectation-maximization methods are based on intensive evaluation of different solutions to select the fittest one. This, combined with the large number of reads present in a modern sequencing experiment, yields to an enormous solution space.

The main body of the program has been developed in R/Bioconductor framework high-level code. This allows an easy and convenient implementation of most of the code and provides a native access to other developed packages. Despite this, the functions used in the inner loop of the genetic algorithm have been implemented in low-level C code for performance purposes. This, combined with a heavy algorithmic work, led to a speed-up of 200X respect initial versions working only with native R functions. Some functions, as the detection of cross-shifts between reads, which require a computational cost of $O(N^2)$, have been replaced to alternative algorithms with a similar outputs with linear cost O(N). Neglecting the technical overhead, the computational cost of a single dynamic is proportional to the number of iterations performed (I), the number of reads analyzed (N) and the population size parameter of the genetic algorithm (P) with an order of magnitude $O(I*(N+P))$. In practical terms, a dynamics of a region of 1000 bases using two datasets with a fold-coverage of 20-50X should take around 5-10 minutes in a 3GHz processor with the default parameters (I=50, P=5000). The memory usage is $O(N*P)$; with default parameters and using a dataset like the one described before we should expect a memory usage of 200-500Mbytes. Once calculated, dynamic's

objects only occupy a few Kbytes. As stated before, different dynamics are based on independent reference objects, allowing an explicit parallelism of the calculations in computer clusters without need to access the source datasets. Furthermore, implicit multicore parallelism is also provided for all computationally intensive functions for executions running in a single machine.

DISCUSSION

Despite nucleosome callers are provide a coarse-grained method for the study of nucleosome positions, we agree with other authors (ref) that the analysis of chromatin variability requires more sophisticated tools which take into account its intrinsic dynamic nature. Approaches based on coverage analysis, including our peak finder nucleR[8], are fast and convenient for most types of downstream studies, but when analyzing the variations inside cell subpopulations the complexity of those analyses grows exponentially. Any method for the analysis of nucleosome variation based on coverage would be able to detect large variations in the coverage profile and attribute them to a unimodal or bimodal changes in the population of reads, but more complex cases changes will remain hidden by the average nature of the coverage peaks.

Third generation sequencing technologies based on single-molecule scanning which does not require PCR amplification are setting the bases for the analysis of variability at single cell resolution. We don't have any doubt that this will be the natural evolution of genomics in the upcoming years, allowing us to rethink the nature of the concept of *fuzziness* in the nucleosome maps and changing it for the one of *dynamics*. In order to understand this new scenario of fine-grain regulation of chromatin and its effects at single cell and locus we need to adapt the existing algorithms for nucleosome analysis. Here we present a bold proposal of a new dimension for nucleosomes studies, based on a shift of paradigm from the study of nucleosome positioning from the peak to the read level.

The limitations of our model based on a expectation maximization are evident, but it marks a first step in the field of read-level nucleosome analysis, which brings this field at the same level than other fields of genomics. The *average* case is not longer representative of the individual variation. As happening in

phenotype characterization by single nucleotide polymorphisms or the prediction of the efficiency to a drug treatment for a given individual patient proposed by personalized medicine, the future of the understanding of nucleosome and chromatin dynamics requires a fine-grain study.

METHODS

Experiment normalization. In the case of experiments with single-end reads, the reads mapping to opposite strands are corrected and a fixed-width nucleosomal fragment is used for dynamics. In both single and paired-end sequencing, over-represented reads attributed to PCR artifacts are removed over the 0.01 threshold[18]. In order to account for different coverage between samples, every occupancy comparison is corrected by a fold-coverage factor. In the operations involving comparison of nucleosome coverage, the short-reads are assumed to have a fixed width, discriminating between the effects of differential digestion from inclusion/eviction.

Detection of regions of interest. Nucleosome Dynamics currently provides three measured for coarse-grained ROIs detection: 1) correlation between coverages, 2) coverage difference per base in window, 3) absolute value of punctual difference. For speed-up, only reads mapping in the positive strand are used in this coarse grained detection of ROIs. This implies that only one mate of the paired-end sequencing is accounted in this step.

ROI retrieval in ENCODE dataset. Only non-blacklisted regions with unambiguously mapping regions obtained from UCSC genome browser (http://genome.ucsc.edu/) were accounted in this study. After this filtering, regions shorter than 1000 bases were discarded. Remaining regions were scanned using the functions provided in Nucleosome Dynamics with running window of 1000 bases with a span of 500 bases. Regions with a mean coverage higher than 20 reads and a correlation less than 0.3 were considered in the study.

Genetic Algorithm. The genetic algorithm features the basic operations involved in this kind of processes: selection of best configurations, crossing, mutation, and insertion of new random configurations. Every configuration is composed by 3 vectors (shifts, variations in width and deletions) where the *i*-th position represents the changes applied read *i.* An additional vector of additional starting positions is kept for insertions. Inserted reads are assumed to have a mean width and are not eligible for other operations. At the start of every execution, a population of configurations is randomly initialized (by default, 5000 instances). During the cycles of the evolution, this population is randomly changed by means

of evolutionary operations and best solution is reported after a given number of iterations (by default 50, execution by default stopped if more than 5 iterations don't provide a fitter solution).

Fitting function for a configuration. The score of a configuration is defined by the product of three independent scores: the coverage and the shift score. Coverage score is the sum of the absolute difference between Ref1' (Ref1 after applying the changes given by the configuration) and Ref2, always normalizing by the ratio of their number of reads. Shift score is the sum of logarithms of the modulus of every shift, causing short shifts weight proportionally more than long ones. An extra penalty for shifts crossing in opposite sense is applied. Optimal scores can be 0 (when Ref1' = Ref2), but multiple theoretical sub-optimal solutions can be achieved depending of the balance between the coverage and the shift score. After testing multiple values, the best subjective balance by both scores was setting the weight of shift scores as the double of coverage score.

Computer resources. Metrics of the calculations presented in this paper are based on executions with the default parameters in an shared memory machine with 32 AMD Opteron cores at 3Ghz. The numbers presented in the paper correspond to a single core execution otherwise stated.

DATA ACCESS

Nucleosome Dynamics can be downloaded through the website http://mmb.irbbarcelona.org/nucdyn. Cell cycle data is available through ENA-SRA website (http://www.ebi.ac.uk/ena) with accession number ERP001689.

ACKNOWLEDGMENTS

We would like to thanks Özgen Deniz for the experimental design of the cell-cycle dataset, Francesc Posas and his collaborators (especially Núria Conde), for advisement regarding Hog1 dataset.

AUTHOR CONTRIBUTION

OF developed the method; OF, MO designed the experiment, analyzed the results and wrote the manuscript.

DISCLOSURE DECLARATION

The authors declare that they have no competing financial interests.

REFERENCES

1. Jiang C, Pugh BF: **Nucleosome positioning and gene regulation: advances through genomics.** *Nature reviews. Genetics* 2009, **10**:161–7210.1038/nrg2522.

2. Mavrich TN, Ioshikhes IP, Venters BJ, Jiang C, Tomsho LP, Qi J, Schuster SC, Albert I, Pugh BF: **A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome.** *Genome research* 2008, **18**:1073–8310.1101/gr.078261.108.

3. Zillner K, Németh A: **Single-molecule, genome-scale analyses of DNA modifications: exposing the epigenome with next-generation technologies.** *Epigenomics* 2012, **4**:403–1410.2217/epi.12.30.

4. Weiner A, Hughes A, Yassour M, Rando OJ, Friedman N: **High-resolution nucleosome mapping reveals transcription-dependent promoter packaging.** *Genome research* 2010, **20**:90–10010.1101/gr.098509.109.

5. Nellore A, Bobkov K, Howe E, Pankov A, Diaz A, Song JS: **NSeq: a multithreaded Java application for finding positioned nucleosomes from sequencing data.** *Frontiers in genetics* 2012, **3**:32010.3389/fgene.2012.00320.

6. Zhang X, Robertson G, Woo S, Hoffman BG, Gottardo R: **Probabilistic inference for nucleosome positioning with MNase-based or sonicated short-read data.** *PloS one* 2012, **7**:e3209510.1371/journal.pone.0032095.

7. Zhang Y, Shin H, Song JS, Lei Y, Liu XS: **Identifying positioned nucleosomes with epigenetic marks in human from ChIP-Seq.** *BMC genomics* 2008, **9**:53710.1186/1471-2164-9-537.

8. Flores O, Orozco M: **nucleR: a package for non-parametric nucleosome positioning.** *Bioinformatics (Oxford, England)* 2011, **27**:2149–215010.1093/bioinformatics/btr345.

9. Zaugg JB, Luscombe NM: **A genomic model of condition-specific nucleosome behaviour explains transcriptional activity in yeast**. *Genome Research* 2011, **22**:gr.124099.111–10.1101/gr.124099.111.

10. Deniz O, Flores O, Martinez PM, Aldea M, Soler-López M, Orozco M: **Nucleosome architecture and plasticity during the cell cycle**. *Submitted*, .

11. Polishko A, Ponts N, Roch KG Le, Lonardi S: **NORMAL: accurate nucleosome positioning using a modified Gaussian mixture model.** *Bioinformatics (Oxford, England)* 2012, **28**:i242–910.1093/bioinformatics/bts206.

12. Schöpflin R, Teif VB, Müller O, Weinberg C, Rippe K, Wedemann G: **Modeling nucleosome position distributions from experimental nucleosome positioning maps.** *Bioinformatics (Oxford, England)* 2013, 10.1093/bioinformatics/btt404.

13. Becker J, Yau C, Hancock JM, Holmes CC: **NucleoFinder: a statistical approach for the detection of nucleosome positions.** *Bioinformatics (Oxford, England)* 2013, **29**:711–610.1093/bioinformatics/bts719.

14. Chen K, Xi Y, Pan X, Li Z, Kaestner K, Tyler J, Dent S, He X, Li W: **DANPOS: dynamic analysis of nucleosome position and occupancy by sequencing.** *Genome research* 2013, **23**:341–5110.1101/gr.142067.112.

15. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, Khatun J, Lajoie BR, Landt SG, Lee B-K, Pauli F, Rosenbloom KR, Sabo P, Safi A, Sanyal A, Shoresh N, Simon JM, Song L, Trinklein ND, Altshuler RC, Birney E, Brown JB, Cheng C, Djebali S, Dong X, Ernst J, et al.: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489**:57–7410.1038/nature11247.

16. Ernst J, Kellis M: **ChromHMM: automating chromatin-state discovery and characterization**. *Nature Methods* 2012, **9**:215–21610.1038/nmeth.1906.

17. Nadal-Ribelles M, Conde N, Flores O, González-Vallinas J, Eyras E, Orozco M, Nadal E de, Posas F: **Hog1 bypasses stress-mediated down-regulation of transcription by RNA polymerase II redistribution and chromatin remodeling.** *Genome biology* 2012, **13**:R10610.1186/gb-2012-13-11-r106.

18. Planet E, Stephan-Otto Attolini C, Reina O, Flores O, Rossell D: **htSeqTools: high-throughput sequencing quality control, processing and visualization in R**. *Bioinformatics (Oxford, England)* 2012, **28**:589–59010.1093/bioinformatics/btr700.

DISCUSSION AND CONCLUSIONS

1. Discussion

1.1. Experimental impact of theoretical DNA mechanics

We present three articles about the connection between intrinsic physical properties of the DNA and the structural and functional properties of chromatin. We found strong evidences that a basal regulatory mechanism underlying the DNA sequence can be correlated with regions of unusual physical properties. Results suggest that intrinsic sequence-dependent properties of DNA play a crucial role in the definition of transcription-active DNA regions. Physical descriptors, combined with chromatin remodelers, transcription factors and epigenetic modifications can therefore define the level of genome activity.

We obtained results suggesting that the impact of physical properties in DNA functionality is related to a modulation of nucleosome positioning around regulatory regions. We don't believe in a universal nucleosome-positioning algorithm able to explain nucleosome architecture in any organism, in fact, our experimental data raise new doubts on the stability of commonly assumed nucleosomal architectures. However, we found strong evidences that regions with unusual physical properties define those regions and, from there, they organize the nucleosome array. Such regions co-localize with regulatory areas.

The role of the DNA methylation regarding nucleosome occupancy is also still not clear. Despite previous biophysical *in vitro* experiments suggesting an increased energetically cost for nucleosomes wrapping in the presence of methylation, recent *in vivo* results positively correlate methylation and nucleosome occupancy. One of the main problems for scientists working in the area is the lack of simple and fast-growing model organisms accounting with DNA methyltransferases on which test the different hypothesis. Biophysical studies *in vitro* and *in silico* of yeast genome methylation strongly suggest an anti-correlation between methylation and nucleosome positioning, which might explain, again using simple physical properties, the connection between DNA functionality and physical properties.

A main conclusion from this thesis is that very basic physical properties of DNA, and perhaps also of RNA (Sciabola *et al.*, 2013), can explain a significant amount of sequence-dependent functional properties of DNA. We believe that DNA contains in itself intrinsic rules that modulate its regulation. Of course, on top of this basal regulatory model, evolution required more complex machinery in order to guarantees a finer control of certain number of highly regulated genes. Therefore, chromatin structure can be the main manifestation by means of which physical properties modulate DNA activity.

1.2. Nucleosome organization *in vivo*

Apart from the validation of theoretical results, we designed new studies with a strong biological meaning. We characterized and quantified the different sources of noise in nucleosomal maps in *S.Cerevisae* and studied the chromatin dynamics during the cell cycle, providing new insights in the understanding of chromosome organization. In all cases we found that many of the experimentally observations had a relationship with simple physical models.

The combination of novel theoretical models and computational tools applied to the study of chromatin has revealed different promoter architectures defined by a synergetic effect between *in vivo* (proteins) and intrinsic (physical properties) factors. During our research of the *in vivo* determinants of nucleosome positioning we acknowledged many variables which we did not take into account in the initial experimental design and which are worthy of an in-deep study. Non-canonical chromatin structures, histone variants, DNA modifications, loci with special characterization such as telomeres or centromers are a new door to open for future research in the group.

1.3. Algorithmic and computational methods

The different methods presented in this thesis are directly linked with the rising of the new high-throughput sequencing and the need of analysis with an increasing level of details. The complexity of the biological systems is growing every day, so it does the available sources of information or performed experiments. This causes a problem commonly known as *paralysis by analysis.*

Every given problem can be faced in many different ways, and it's impossible to cover them all. The need of reliable, robust data is linked with the will of finding general and reproducible results across different experiments, and those requirements are sometimes opposite. As presented in this thesis,

current technologies provide a blur, averaged, vision of what is happening in a cell population, and the detailed knowledge of specific mechanisms needs a clear vision of what is happening inside a single cell. Deriving robust information from the fuzzy population, averaged data is one of the largest challenges for bioinformatics in the next years. Our contribution in this sense was to provide a fast and reliable tool for the mining of nucleosome calls, nucleR, which eliminates the need of modeling and enabled the classification of different chromatin architectures. In our collaboration with IRB biostatistical unit, we faced the need of high quality experimental data by helping in the development of htSeqTools. We also enhanced the possibilities of automatic annotation within the R framework providing new interfaces to on-line repositories in DASiR. Finally, a challenging new method for the deep-analysis of chromatin variability between experiments is presented in Nucleosome Dynamics package, accounting for the growing need of automatic, dynamic and low-level processing of genomic data.

The integration of those tools with existing or novel analysis platforms could allow new possibilities of analysis to the scientific community in a not fully explored field: next-generation-sequencing analysis of chromatin structure at high resolution level.

2. Subproject Conclusions

2.1. Physical properties of naked DNA influence nucleosome positioning and correlate with transcription start and termination sites in yeast

- MNase has a significant affinity for cleavage between AT basepairs even in the absence of histones (naked DNA).
- Key nucleosome depleted, gene-regulatory regions such as TSS or TTS are signaled by a differential MNase digestion pattern in naked DNA.
- MNase degradation profiles of nucleosomal and naked DNA correlate with unusual stiffness properties.
- The physical properties explaining the increased MNase affinity also describe a higher energy potential for nucleosome formation in the affected areas.
- Special characterization of DNA physical properties correlate with nucleosome positioning and signal gene regulatory regions in yeast and other organisms.

2.2. Impact of Methylation on the Physical Properties of DNA

- Physical characterization of dinucleotide steps can be altered accordingly the neighboring bases, featuring a bimodal behavior in for some parameter-basepair pairs.
- Despite d(CpG) steps are very flexible, long d(CpG) tracks feature an average flexibility.
- Methylation of d(CpG) steps translates into a higher curvature and stiffness of the basepair.
- Neglecting the effect of external factors, the d(CpG) methylation increases the deformation energy required to wrap DNA around nucleosome and can affect the nucleosome architecture in regulatory regions, changing phasing or even displacing nucleosomes.

2.3. Unraveling the hidden DNA structural/physical code provides novel insights on promoter location

- Comprehensive analysis of ProStar (Goñi *et al.*, 2007) false-positive predicted (according to 2007-state of knowledge) promoters in human enabled the discovery of novel functional promoters in human.
- CAGE and RNA-Seq analysis provided evidences of unidirectional transcript activity.

- High-confidence ProStar predicted regions, sharing a defined pattern of physical features, truly behave like physiologically active TSSs, reinforcing connection between regulatory regions and unusual regions of DNA.

2.4. Nucleosome architecture and plasticity along cell cycle

- Chromatin varies along cell cycle. S phase displays differential chromatin structures.
- At individual gene level, the change in the nucleosome positioning pattern around TSS is clearer in cell cycle dependent genes.

2.5. Fuzziness and noise in nucleosomal architecture

- Technical and biological noise can lead to misleading nucleosome map comparison. Variation in the maps available in literature could have been overinterpreted.
- Single-end sequencing is in general less robust than paired-end sequencing, but in the later, underdigested dinucleosome fragments can bias the normalized coverage signal.
- Inter-replica variations are not negligible as nucleosome positioning is revealed as intrinsically plastic and dynamic.
- Cell-cycle synchronized samples in G1 appear as less noisy and are more reproducible than other comparable asynchronous experiments.
- Excessive MNase digestion can partially degrade some well-positioned nucleosomes, meanwhile an under-digestion can lead to large DNA fragments difficult to sequence increasing the observed non covered regions.
- Basal lowly expressed genes display better defined nucleosome organization than highly expressed genes.
- Two main classes of nucleosome architectures exist on promoters: open and closed NFR.
- There is a direct relationship between the open-close profile of NFR and the deformation energy derived from physical descriptors of DNA. A synergetic effect with transcription factors can explain most of the observed architectures.
- If a large energetic barrier exists in the NFR (for example in the open NFR), nucleosomes are then phased periodically from this point.

2.6. Hog1 bypasses stress-mediated down-regulation of transcription by RNA polymerase II redistribution and chromatin remodeling

- Environmental stress causes major changes of the RNA Pol II towards stress-responsive genes, meanwhile housekeeping genes are down-regulated
- RNA Pol II and III re-localization require the stress-activated protein kinase Hog1
- A strong chromatin remodeling process is observed in stress-responsive genes upon stress. This process is dependent of Hog1.
- Hog1 serves to bypass the general down-regulation of gene expression that occurs in response to osmostress by targeting RNA Pol II machinery and by inducing chromatin remodeling at stress-responsive loci.

2.7. Modeling genome-wide kinetics of endo/exonuclease digestion

- MNase exo- and endo-nuclease activity in *C.Elegans* follows the same principles as that described previously by our group in *S.Cerevisiae.*
- The sequence affinity of MNase is more remarkable in short digestions.
- MNase-seq profiles obtained after pooling samples with different digestion times reveal preferentially short, long or dynamic nucleosomes.
- Loci more sensible to short MNase digestions are located in high deformation energy areas, meanwhile resistant loci show an average energy profile.

2.8. nucleR: a package for non-parametric nucleosome positioning

- nucleR features a basic pre-processing of MNase-seq experiments and Tiling Microarrays.
- A non-parametric, fast and efficient algorithm for nucleosome calling was implemented using a FFT component knock-out as a noise-gate for the coverage profiles

2.9. htSeqTools: high-throughput sequencing quality control, processing and visualization in R

- htSeqTools allows different quality control procedures for high-throughput sequencing experiments. It includes detection of outliers and biases, inefficient immuno-precipitation and over-amplification artifacts.

- Other features include the *de novo* identification of read-rich genomic regions and visualization of the location and coverage of genomic region lists.

2.10. DASiR: Programmatic data retrieval from DAS servers in R

- DASiR presents a convenient interface to access Distributed Annotation System servers within R framework.

2.11. Dynamic analysis of nucleosome positioning at read level

- Nucleosome Dynamics allows the calculation of changes between two nucleosome maps at the read level
- The program describes these dynamics in terms of shifts, insertions, deletions and changes in the read size.
- Use cases include automatic genome annotation, quantification of chromatin dynamics between two experiments and the graphical representation of the dynamics.

3. General conclusions

- Nucleosomes are highly-dynamic entities and the study of their positioning requires systematic experimental design and very accurate bioinformatics analysis.
- Physical properties of DNA are a key player to explain regulatory and functional properties of DNA. They must be taken into account in genomic analysis as any other of the factors usually used to determine the causes of a given biological phenomena.
- Nucleosome architecture is fuzzier than anticipated, but the connection between nucleosome arrangements, physical properties and DNA functionality is an emerging theory.
- The interplay between nucleosomes, DNA modifications, DNA mechanics and abnormal chromatin structures are a key subject in the understanding of basic biological processes as well as the advance towards the definition, and therefore the cure, of different diseases.

RESUM (CATALÀ)

1. Introducció

En els organismes eucariotes, l'ADN es troba compactat dins el nucli de cada cèl·lula gràcies a l'estructura en forma de fibres de cromatina proporcionada pels nucleosomes. Els nucleosomes actuen per tant com un element tan estructural com regulador del genoma, exposant o limitant l'accés de la seqüència primària al solvent (Richmond and Davey, 2003).

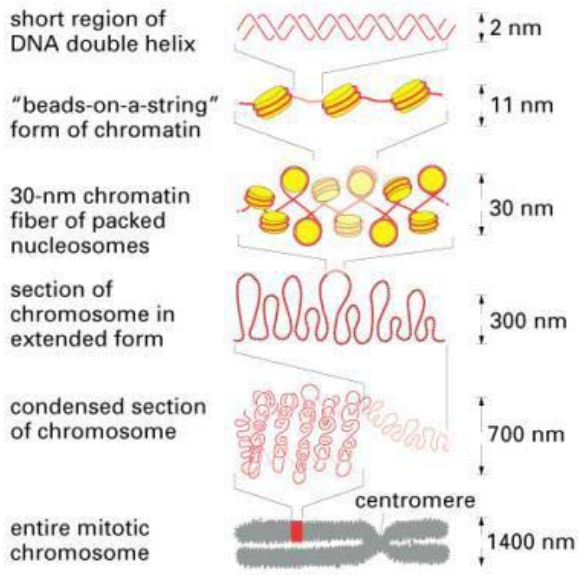


Figura 1. Els diferents nivells de compactació de la cromatina, des de la doble hèlix de l'ADN (superior) fins al cromosoma (inferior) (Alberts *et al.*, 2004) (Adaptació)

Consistentment amb el seu rol regulador, els nucleosomes no es distribueixen de forma aleatòria al llarg del genoma. En certes regions, normalment amb activitat funcional, podem observar un posicionament comú en diverses cèl·lules, originant una visió global en fase que dona nom als nucleosomes *ben posicionats*. Altres regions amb un posicionament més dinàmic reben el nom de nucleosomes *deslocalitzats*. Anomenarem *cobertura* al nombre de nucleosomes que ocupa un o altre tipus en una certa regió del genoma, mentre que les regions no ocupades poden ser tant l'espai d'unió entre dos nucleosomes (ADN d'enllaç), normalment al voltant d'uns 20 parells de bases, o regions més amplies que per una o altra raó es troben lliures de nucleosomes (Mavrich, Ioshikhes,

et al., 2008). La combinació d'aquestes propietats es pot veure clarament en els llocs d'inici de transcripció (TSS, per les seves inicials en anglès), on la representació típica mostraria una regió lliure de nucleosomes just abans del TSS, amb nucleosomes ben posicionats a banda i banda que anirien perdent la sincronia a mesura que ens allunyem d'aquest lloc.

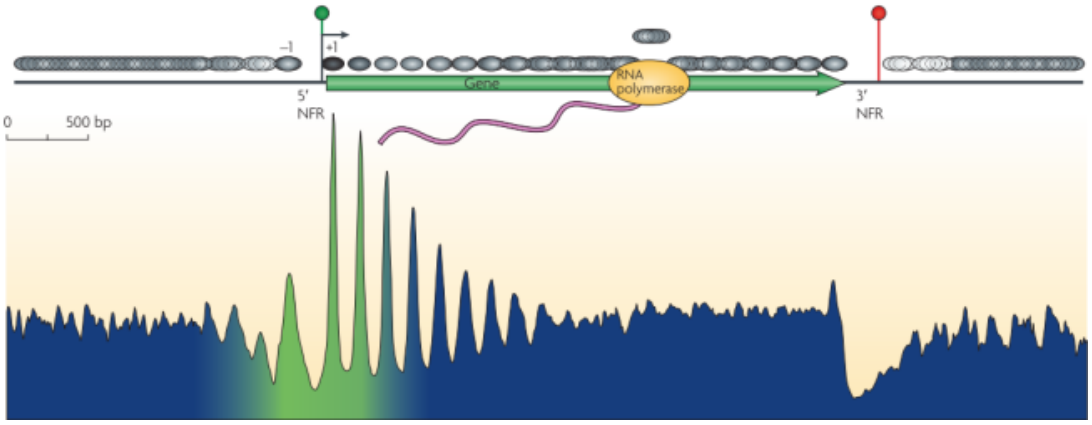


Figura 2. Distribució esperada dels nucleosomes en llevat. La representació superior representaria una cadena de nucleosomes (cercles grisos) al voltant del cos d'un gen donat (fletxa verda). La regió lliure de nucleosomes (NFR) apareix als extrems 5' i 3' dels gens. A la part inferior, es representa la cobertura de nucleosomes. Grans acumulacions de nucleosomes en fase apareixen com a pics alts i prims, mentre que un senyal més difós marcaria una pèrdua de la sincronia del seu posicionament entre diferents cèl·lules. (Jiang and Pugh, 2009) (Adaptació)

Des de l'aparició de les tecnologies genòmiques d'alt rendiment i resolució durant la darrera dècada, l'estudi del posicionament dels nucleosomes en el genoma ha estat un tema àmpliament estudiat. Diversos estudis mostren unes preferències rotacionals en la col·locació de les bases de l'ADN al voltant del nucli d'histones, donant lloc a la teoria de que el posicionament dels nucleosomes podria estar regulat per la pròpia seqüència subjacent (Segal *et al.*, 2006). Malgrat tot, el posicionament *in vivo* dels nucleosomes és altament dinàmic i difícil de predir, per contra del que passa amb experiments de reconstitució *in vitro*. Diversos factors epigenètics, com la metilació de l'ADN o la modificació de les histones, han mostrat tenir un paper regulador en l'expressió genètica estar directament relacionats amb el posicionament dels nucleosomes (Segal and Widom, 2009). Estudis recents també demostren la funció dels remodeladors de cromatina en l'expressió genètica, donant lloc a un conglomerat d'elements que enriqueixen el rol dels nucleosome més enllà del que es pensava fa uns anys (Segal and Widom, 2009).

Al llarg d'aquesta tesi es presenten diversos treballs com a fruit de l'aplicació de les tecnologies experimentals i computacionals més recents relacionats amb l'estudi dels nucleosomes i la cromatina. Si bé la majoria d'aquests treballs es poden englobar sota el nom que dona lloc a aquesta tesi, Posicionament de nucleosomes en cromatina mitjançant tècniques bioinformàtiques, una visió més detallada donarà lloc a tres camps de coneixement diferent.

En primer lloc farem una aproximació teòrica a les propietats físiques de l'ADN i el seu impacte en les funcions *in vivo*. L'ADN, com a un polímer semi flexible que és, té associades unes propietats mecàniques que permeten que certes seqüències siguin més o menys mal·leables. Això pot afectar a la seva propietat d'embolcallar el nucli d'histones en la conformació del nucleosoma (Xu and Olson, 2010). Diversos treballs han mesurat les propietats físiques a nivell teòric d'una cadena d'ADN donada a partir de models elàstics derivats de les diferents constants de flexibilitat a nivell de parell de base (Pérez *et al.*, 2008; Lankas *et al.*, 2003)(Figura 3).

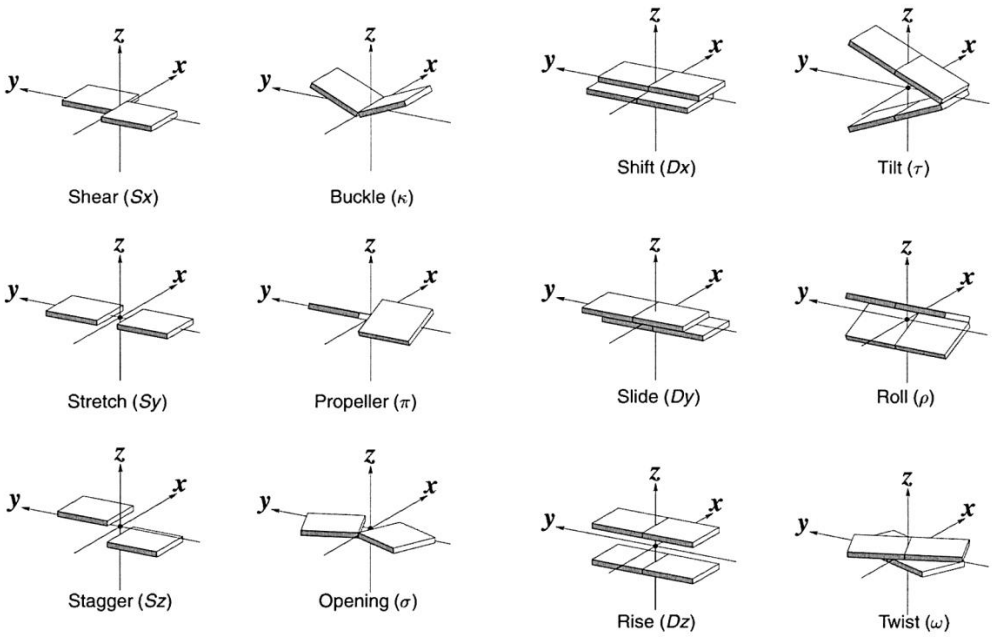


Figura 3. Paràmetres helicoïdals a nivell de base (dues primeres columnes) o a nivell de parell de base (últimes dos columnes). Es mostren 3 translacions (primera columna de cada grup) i tres rotacions (segona columna de cada grup) per cadascuna de les tres dimensions de l'espai. (Lu and Olson, 2003)

Nosaltres mostrem en diversos treballs diferents relacions entre aquestes propietats i un model mesoscòpics per calcular l'energia de deformació necessària per passar d'una conformació relaxada d'ADN a les 1.6 voltes necessàries per formar un nucleosoma. Aquests treballs són: *Physical proper-*

ties of naked DNA influence nucleosome positioning and correlate with transcription start and termination sites in yeast (Les propietats físiques del ADN nu influencien el posicionament dels nucleosomes i es correlacionen amb els llocs d'inici i final de transcripció en llevat), *Impact of Methylation on the Physical Properties of DNA* (Impacte de la metilació en les propietats físiques de l'ADN) i *Unraveling the hidden DNA structural/physical code provides novel insights on promoter location* (Desentranyar el codi físic/estructural ocult en l'ADN proporciona nous coneixements sobre la localització de promotors).

En segon lloc, trobarem diversos estudis de naturalesa principalment experimental. En els darrers anys s'han publicat diferents mapes de nucleosomes que han permès la millor comprensió de l'organització dels nucleosomes *in vivo* (Yuan *et al.*, 2005; Lee *et al.*, 2007; Schones *et al.*, 2008; Li *et al.*, 2011). En general, aquests mapes s'obtenen de digerir la cromatina amb nucleasa micrococcal (MNase) obtenint així fragments d'ADN nucleosomal que posteriorment es poden seqüenciar mitjançant tècniques d'alt rendiment com microarrays o seqüenciadors d'ADN .

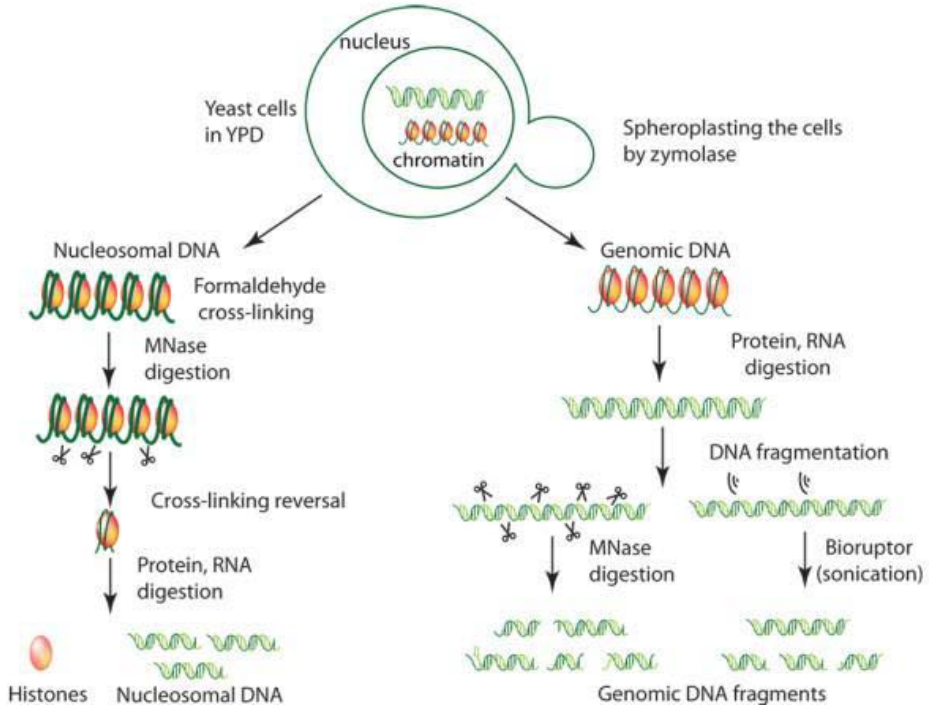


Figura 4. Esquema del protocol per l'obtenció de l'ADN nucleosomal i genòmic (control). (Deniz *et al.*, 2011)

En el nostre treball hem estudiat la dinàmica dels nucleosomes al llarg del cicle cel·lular en el treball *Nucleosome architecture and plasticity along cell cycle* (Arquitectura i plasticitat dels nucleosomes al llarg del cicle cel·lular) així com les possibles fonts d'imprecisió en la mesura de dites dinàmiques en l'article *Fuzziness and noise in nucleosomal architecture* (Imprecisió i soroll en l'arquitectura nucleosomal). També s'inclouen en aquesta secció dues col·laboracions, *Hog1 bypasses stress-mediated down-regulation of transcription by RNA polymerase II redistribution and chromatin remodeling* (Hog1 evita la caiguda de la transcripció per estrès mitjançant la redistribució de la ARN polimerasa II i el remodelatge de la cromatina) i un estudi sobre l'activitat exo-/endo-nucleasa de la nucleasa micrococcal en c.elegans realitzat durant la meva estada en el grup del Prof. Jason Lieb a la Universitat de Nord Carolina l'estiu del 2013.

Tots aquests anàlisis no serien possible si no contéssim amb determinades eines pel processament massiu de les dades generades des del laboratori experimental. Un primer conjunt d'eines inclouri-en models per tal d'inferir el posicionament discret dels nucleosomes a partir de les dades de seqüenciació massiva.

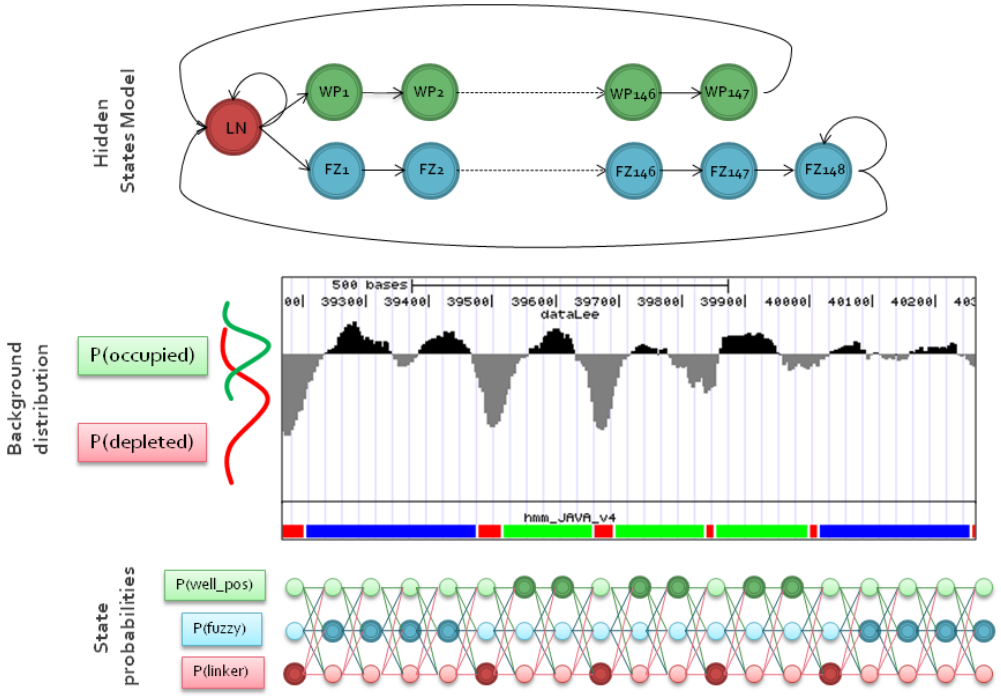


Figura 5. Implementació pròpia d'un Model de Markov Ocult (HMM) pel posicionament de nucleosomes a partir de dades de microarray inspirat en un treballs previs (Yassour *et al.*, 2008).

Si bé aquest procés requeria d'una potent maquinària estadística per tal de modelar les dades informatives en mig del soroll, nosaltres vam proposar una nova manera de fer-ho a partir d'un filtratge de les dades seguit d'un simple procés de reconeixement de pic amb la eina *nucleR: a package for non-parametric nucleosome positioning* (nucleR: un paquet pel posicionament no paramètric de nucleosomes).

Evidentment, estudis més sofisticats requereixen un nivell de detall superior al que proporcionen rangs de nucleosomes discrets, especialment en treballar amb dinàmiques entre experiments (Chen *et al.*, 2013). En aquest sentit, nosaltres hem proposat l'eina de Nucleosome Dynamics, presentada a l'article *Dynamic analysis of nucleosome positioning at read level* (Anàlisis dinàmic del posicionament de nucleosomes al nivell de lectures de seqüències).

Finalment, i com a part de col·laboracions amb altres experts, també em desenvolupat o contribuït en el desenvolupament de noves eines per un processament més eficient i de major qualitat de les dades en un entorn bioinformàtic. Al respecte, es poden consultar els treballs *htSeqTools: high-throughput sequencing quality control, processing and visualization in R* (htSeqTools: control de qualitat, processament i visualització d'experiments de seqüenciació d'alt rendiment en R) i *DASiR: Programmatic data retrieval from DAS servers in R* (DASiR: Recuperació programàtica de la informació des de servidors DAS en R)

2. Objectius

2.1. Impacte experimental de la mecànica teòrica de l'ADN

- Investigar la relació entre la flexibilitat intrínseca de diferents seqüències i regions reguladores o mapes de nucleosomes *in vivo*.
- Ajustar els models mesoscòpics existents de flexibilitat de l'ADN per suportar bases metilades i l'anàlisi de genomes sencers
- Proporcionar suport bioinformàtic al laboratori experimental en la validació de resultats teòrics

2.2. Organització dels nucleosomes *in vivo*

- Processar dades de MNasa-seq per tal d'obtenir els nostres propis mapes de nucleosomes
- Realitzar controls de qualitat i analitzar possibles biaixos en experiments de MNasa-seq
- Estudiar l'impacte en la cromatina de l'expressió diferencial de gens i la replicació de l'ADN durant el cicle cel·lular en llevat.
- Establir col·laboracions amb grups experimentals amb una llarga trajectòria per la recerca conjunta de l'organització i regulació de la cromatina, proporcionant-los la nostra experiència i recursos computacionals

2.3. Algoritmes i mètodes computacionals

- Desenvolupar un predictor del posicionament dels nucleosomes basat en descriptors físics de l'ADN.
- Establir un protocol pel processament i anàlisis de dades de xips d'ADN i experiments de seqüenciació, especialment centrat en mapes de nucleosomes obtinguts d'experiments MNasa-seq
- Revisar mètodes existents apropiats pels nostres anàlisis i desenvolupar-ne de nous si fos necessari.
- Adaptar algoritmes o mètodes existents per aprofitar els supercomputadors d'alt rendiment disponibles en el grup.

3. Resum de les publicacions (abstractes)

3.1. Physical properties of naked DNA influence nucleosome positioning and correlate with transcription start and termination sites in yeast

Les propietats físiques del ADN nu influencien el posicionament dels nucleosomes i es correlacionen amb els llocs d'inici i final de transcripció en llevat

Antecedents: En els organismes eucariotes l'ADN s'empaqueta en forma de cromatina, on la major part de l'ADN està troba embolcallat en nucleosomes. La compactació de l'ADN i el posicionament dels nucleosomes tenen unes implicacions funcionals clares, ja que modulen l'accessibilitat de les proteïnes reguladores a determinades regions genòmiques. Malgrat la investigació intensiva en aquesta àrea, les regles que defineixen el posicionament de nucleosomes i la ubicació de les regions reguladores de l'ADN segueixen sent de difícil comprensió.

Resultats: Mostres d'ADN nu (sense histones) i ADN nucleosomal de llevat van ser digerides amb nucleasa micrococal i posteriorment es va seqüenciar el genoma. Es van determinar les preferències de tall d'aquest enzima per ambdues mostres. La integració dels perfils d'ocupació fruït de la seqüenciació es van combinar amb descriptors conformacionals de l'ADN derivats de simulacions atomístiques de dinàmica molecular, permetent-nos extreure les propietats físiques de l'ADN a nivell genòmic i correlacionar-los amb l'estructura de la cromatina i la regulació genòmica. Hem trobat que l'estructura local de l'ADN al voltant de regions reguladores és extraordinàriament flexible i mostra un patró singular de posicionament dels nucleosomes. Descriptors físics *ab initio* derivats de les dinàmiques moleculars es poden utilitzar per desenvolupar un mètode computacional que prediu amb precisió regions enriquides o empobrides de nucleosomes.

Conclusions: Les nostres anàlisis experimentals i computacionals demostren una clara correlació entre les propietats físiques dependents de la seqüència d'ADN i diverses senyals reguladores de l'estructura de la cromatina. Aquests resultats demostren que, almenys en el llevat, el posicionament de nucleosomes al voltant del lloc d'inici i final de transcripció esta fortament relacionat amb les propietats físiques de l'ADN, podent definir un mecanisme de regulació basal de l'expressió gènica.

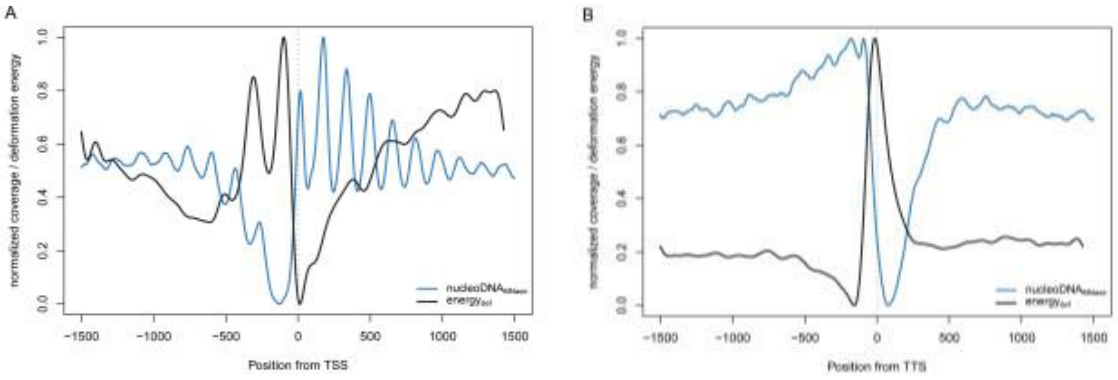


Figura 6. Perfils de posicionament experimental de nucleosomes (blau) I perfils de l'energia de deformació predita (negre) en les regions del TSS (A) I TTS (B)

3.2. Impact of Methylation on the Physical Properties of DNA

Impacte de la Metilació en les propietats físiques de l'ADN

Cada cop hi ha més evidències de la presència d'un codi alternatiu imprès en el genoma que podrien contribuir a la regulació de l'expressió gènica a través d'un mecanisme indirecte de lectura. En mamífers, els components d'aquest mecanisme de regulació a baix nivell inclouen l'estructura de la cromatina i les signatures epigenètiques, on els parells de nucleòtids d(CpG) en són actors clau. En aquest treball informem sobre un estudi experimental i teòric exhaustiu dels parells d(CpG) que proporciona una descripció detallada de les característiques físiques i una quantificació de l'impacte de la metilació de citosines en les propietats físiques de l'ADN. Hem observat que la metilació canvia aquests propietats físiques en parells d(CpG), afectant dramàticament segments enriquits en CpG, com ara illes CpG. Hem demostrat que la metilació redueix l'afinitat de l'ADN a formar estructures afins a nucleosomes, podent afectar el posicionament d'aquests al voltant dels llocs d'inici de transcripció. En general, els nostres resultats suggereixen un mecanisme pel qual les propietats físiques bàsiques de la fibra d'ADN poden explicar part dels mecanismes de regulació epigenètics cel·lulars.

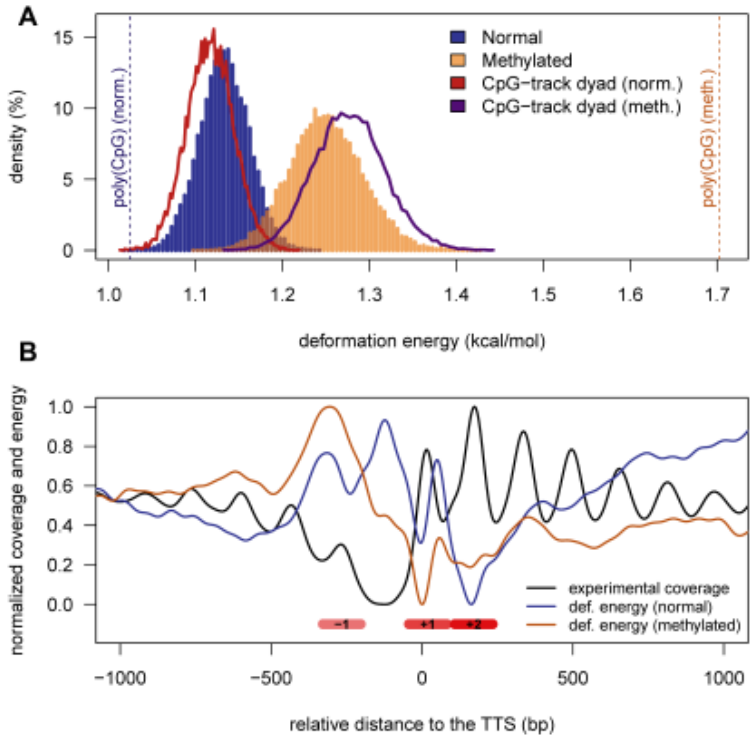


Figura 7. Efectes teòrics de la metilació de l'ADN en la formació de nucleosomes. A) Distribució d'energia de deformació per ADN normal i metilat en un conjunt aleatori de seqüències genòmiques. B) Cobertura experimental de nucleosomes (negre) comparada amb l'energia de deformació mitjana en la regió, assumint un estat normal de les bases CpG (blau) o metilat (taronja).

3.3. Unraveling the hidden DNA structural/physical code provides novel insights on promoter location

Desentranyar el codi físic/estructural ocult en l'ADN proporciona nous coneixements sobre la localització de promotors

Tot i que el reconeixement de motius d'ADN en regions promotores s'ha considerat tradicionalment com un element regulador crític en la transcripció, la ubicació dels promotors i en particular els llocs d'inici de transcripció, encara continua sent un repte. A continuació es presenta una anàlisi exhaustiva de possibles seqüències promotores no anotades al llarg del genoma humà definides per diferents propietats físiques de l'ADN considerades en el nostre algorisme computacional ProStar. Un mostreig representatiu de les regions predites es va validar extensament i analitzar experimentalment. Curiosament, la gran majoria van demostrar ser seqüències transcripcionalment actives tot i la manca de motius de seqüències específics. Això indica que la senyalització física és de fet

capaç de predir l'activitat d'un promotor més enllà dels mètodes de predicció convencionals. D'altra banda, regions altament transcrites mostren característiques en la cromatina típicament associades als promotors dels gens constitutius. Els nostres resultats permeten redefinir el que entenem com a signatura d'un promotor i analitzar la diversitat, la conservació evolutiva i regulació dinàmica dels diverses regions reguladores fonamentals en humans. D'altra banda, el present estudi dóna suport a la hipòtesi d'un mecanisme evolutiu de regulació indirecta codificada en les propietats físiques intrínseques de l'ADN que poden contribuir a la complexitat de la regulació genètica en humans.

3.4. Fuzziness and noise in nucleosomal architecture

Imprecisió i soroll en l'arquitectura nucleosomal

L'organització dels nucleosomes juga un paper clau en la regulació de l'expressió gènica. A pesar d'això, malgrat els notables avenços en la precisió dels mapes de nucleosomes, encara hi ha severes discrepàncies sobre el posicionament individual dels nucleosomes i com això influencia la regulació gènica. La variabilitat entre diferents mapes de nucleosomes, que impedeix l'anàlisi detallat del seu posicionament, te el seu origen en diverses fonts. Hem inspeccionat detalladament els diferents factors extrínsecs que poden induir aquesta diversitat comparant diferents mapes de nucleosomes a partir d'experiments de MNase-Seq generats sota diferents condicions. Addicionalment, també hem explorat la variació originada per factors intrínsecs a la dinàmica de la cromatina generant mapes de poblacions de cèl·lules sincronitzades pel que fa el cicle cel·lular. Analitzar aquestes dades ens ha permès mesurar l'efecte del soroll en l'ocupació i posicionament dels nucleosomes així com ens ha proporcionat una nova visió sobre els seus determinants subjacents. Presentem, doncs, una aproximació sistemàtica que pot guiar la estandardització dels experiments de MNase-Seq per tal de generar patrons de nucleosomes reproduïbles.

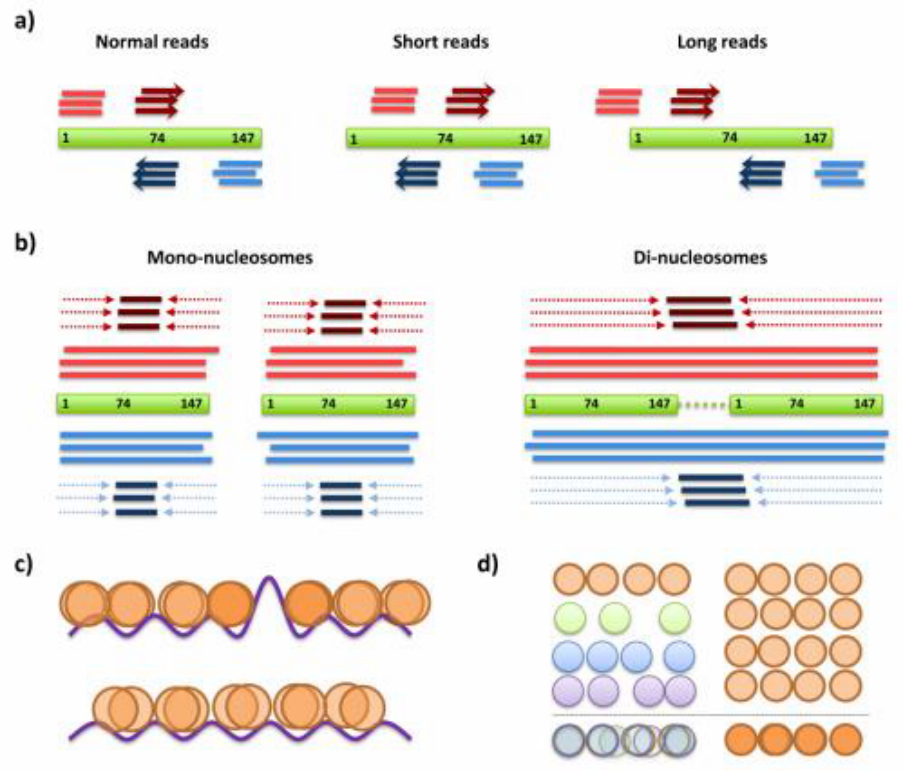


Figura 8. Fonts de dispersió i soroll en els mapes de nucleosomes. a) Desalineament de les lectures en *single-end sequencing* b) presència de di-nucleosomes en *paired-end sequencing,* c) barreres d'energia que actuen com un element posicionador i d) cèl·lules en estat de creixement heterogeni enlloc de treballar amb una població de cèl·lules sincronitzades.

3.5. Hog1 bypasses stress-mediated down-regulation of transcription by RNA polymerase II redistribution and chromatin remodeling

Hog1 evita la caiguda de la transcripció per estrès mitjançant la redistribució de la ARN polimerasa II i el remodelatge de la cromatina

Antecedents: Les cèl·lules es veuen sotmeten a canvis dràstics de l'expressió gènica degut canvis ambientals. Aquest estrès provoca una regulació a la baixa de l'expressió gènica general, juntament amb la inducció d'un conjunt de gens de resposta a estrès. La proteïna quinasa de resposta a l'estrès Hog1, relacionada amb p38, és un important regulador de la transcripció en resposta a canvis en l'osmolaritat en el llevat.

Resultats: Estudis de localització en el genoma de l'ARN polimerasa II (ARN Pol II) i Hog1 mostren que l'estrès indueix importants canvis en la localització de l'ARN Pol II, amb un increment en

l'expressió de gens de resposta a estrès en comparació amb gens constitutius. La relocalització de l'ARN Pol II requereix Hog1, que també es troba localitzada en *loci* de resposta a estrès. A més dels gens amb unió a ARN Pol II, Hog1 també co-localitza en gens amb unió a RNA polimerasa III, cosa que apunta a un paper més ampli del control transcripcional exercit per Hog1 que el previst inicialment. Curiosament, associacions més fortes de Hog1 amb gens de resposta a estrès es correlacionen amb la remodelació de la cromatina i un augment de l'expressió gènica. Cal destacar que l'anàlisi de MNasa-seq mostra que encara que l'estructura de la cromatina no s'altera significativament a gran escala en resposta a l'estrès, hi ha un pronunciat efecte degut a la remodelació de la cromatina en gens que mostren una associació amb Hog1.

Conclusió: Hog1 serveix per evitar la baixada en la regulació general de l'expressió gènica que es produeix en resposta a osmoestrès, i ho fa tant a través de la maquinària de l'ARN Pol II com mitjançant la inducció de remodelació de la cromatina en els *loci* de resposta a estrès.

3.6. nucleR: a package for non-parametric nucleosome positioning

nucleR: un paquet pel posicionament no paramètric de nucleosomes

nucleR és un paquet per R/Bioconductor que permet un reconeixement ràpid i flexible del posicionament de nucleosomes provinents d'experiments de seqüenciació o de xips d'ADN. Aquest programari es troba integrat en el conjunt de paquets de R per l'anàlisi genòmic d'alt rendiment i permet la visualització *in situ*, així com l'exportació dels resultats als formats més comuns suportats pels visualitzadors de genoma.

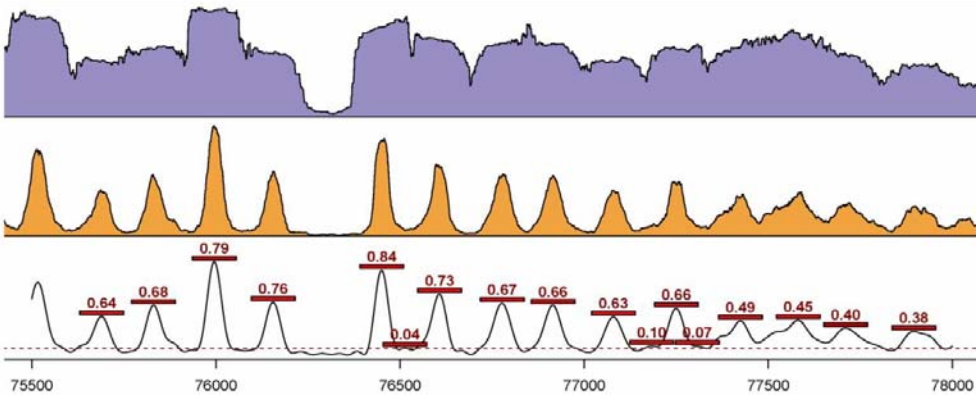


Figura 9. Exemple del funcionament de nucleR. De dalt a baix, el perfil de cobertura en cru, el perfil preprocessat I el perfil amb el soroll filtrat sobre el que es busquen els pics.

3.7. htSeqTools: high-throughput sequencing quality control, processing and visualization in R

htSeqTools: control de qualitat, processament i visualització d'experiments de seqüenciació d'alt rendiment en R

Proporcionem un paquet de Bioconductor amb eines per l'avaluació de la qualitat, processament i visualització de dades de seqüenciació d'alt rendiment, amb èmfasi en els estudis de ChIP-seq i RNA-seq. S'inclou la detecció de valors atípics, la correcció de biaixos, immunoprecipitació ineficient i artefactes deguts a la sobreamplificació, identificació *de novo* de regions enriquides en lectures de seqüència i la visualització de la ubicació i la cobertura de regions genòmiques.

3.8. DASiR: Programmatic data retrieval from DAS servers in R

DASiR: Recuperació programàtica de la informació des de servidors DAS en R

El Sistema d'Anotació Distribuïda (DAS, per les seves inicials en anglès) és un protocol per l'intercanvi d'informació entre un client i un servidor. És àmpliament emprat en bioinformàtica i els repositoris de dades més importants solen incloure un servidor DAS paral·lelament a la seva interfície principal. UCSC, Ensembl o Uniprot en son alguns exemples. El paquet DASiR proporciona una còmoda interfície R-DAS permetent l'accés programàtic als servidors DAS disponibles en la xarxa. Suporta les principals característiques del protocol DAS 1.6 permetent l'accés a enormes quantitats d'informació biològica des de R. DAS s'implementa sobre els protocols XML i HTTP, pel que els requeriments pel seu desplegaments son significativament menors que els emprats per altres servidors, com MySQL o Biomart. Es pot trobar una llista navegable amb mes de 1500 servidors DAS en la direcció http://www.dasregistry.org

3.9. Dynamic analysis of nucleosome positioning at read level

Anàlisis dinàmic del posicionament de nucleosomes al nivell de lectures de seqüències

L'estudi dels mapes de nucleosomes en les últimes dècades ha posat de manifest que el paper regulador de la cromatina regions reguladores va més enllà de la simple oclusió de l'ADN segons l'estat transcripcional del gen. Treballs recents assenyalen cap a la necessitat d'estudiar la variabilitat dels nucleosomes tenint en compte els seus efectes específics sobre subpoblacions de cèl·lules en lloc de capturar només les seves tendències generals. Paral·lelament a això, els avenços en la tecnologia

de seqüenciació i el creixent poder computacional disponible permeten formular noves preguntes que eren tècnicament impossibles fins ara. En aquest treball presentem *Nucleosome Dynamics*, un algoritme que permet l'anàlisi comparativa del posicionament dels nucleosomes a nivell d'una lectura de seqüència. Això permet la caracterització dels canvis observables a nivell de cobertura, així com proporciona una explicació de la seva causa en un context dinàmic. L'anotació a nivell de lectura de seqüència permet un anàlisi d'alta resolució sense precedents fins ara, incloent la detecció d'inclusions, eviccions, desplaçaments o canvis en la digestió de la cromatina. S'exemplifica la seva aplicació en tres casos d'ús diferents: la caracterització de les diferències a nivell de cromatina entre dues línies cel·lulars humanes, la quantificació dels canvis soferts entre cèl·lules de llevat en estat natural i amb mutació en HOG1 com a resposta a estrès i la representació dels canvis durant les diferents etapes del cicle cel·lular en llevat.

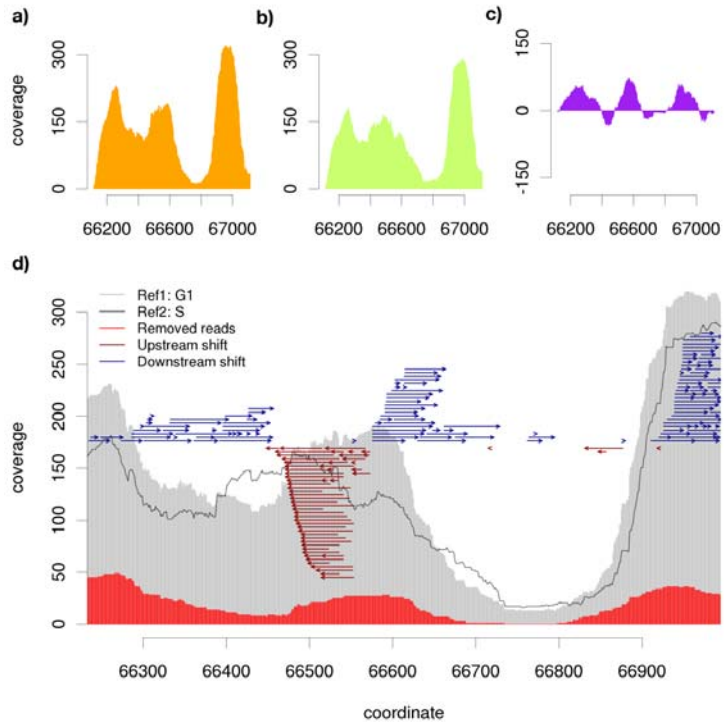


Figura 10. a, b) Perfils de cobertura de nucleosomes per la referència 1 i 2, corresponents una regió donada del genoma del llevat. c) Diferència base a base dels dos perfils. Això permet una quantificació però no una explicació dels canvis en el perfils. d) Anotació a nivell de lectura proporcionada per Nucleosome Dynamics. Concretament es mostra una dispersió de nucleosomes cap en el centre del gràfic.

3.10. Treballs sense publicació en el moment del dipòsit

Nucleosome architecture and plasticity along cell cycle

Arquitectura i plasticitat dels nucleosomes al llarg del cicle cel·lular

Per tal d'estudiar el posicionament dinàmic dels nucleosomes al llarg del cicle cel·lular, hem obtingut diferents mapes de nucleosomes en diferents etapes de creixement (G1, S, G2 i M). Els resultats obtinguts mostren una desorganització més accentuada en les etapes amb major canvi de la cromatina, especialment durant l'etapa de replicació de l'ADN en la fase S. Els canvis al nivell individual mostren com gens amb una expressió diferencial al llarg del cicle mostren perfils variants en la seva arquitectura al voltant del inici de transcripció.

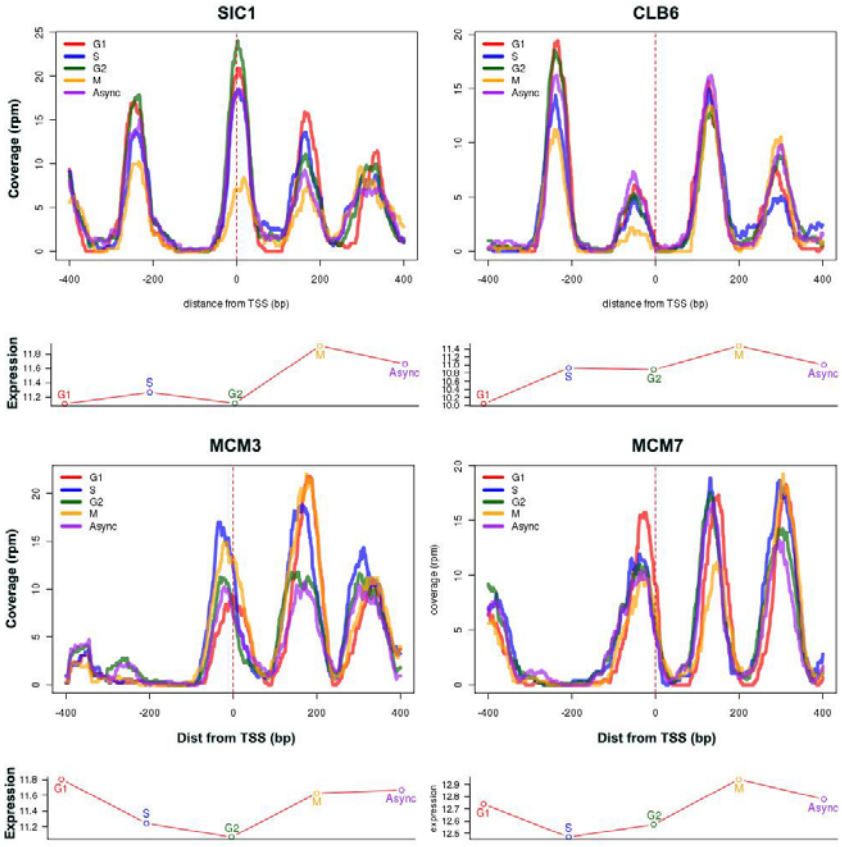


Figura 11. Perfils individuals del posicionament de nucleosomes (a dalt) I perfils d'expressió dels gens SIC1, CLB6, MCM3 I MCM7, involucrats en el cicle cel·lular.

Modeling genome-wide kinetics of endo/exonuclease digestion

Modelatge a nivell de genoma de l'activitat de digestió endo/exonucleasa

Durant l'estiu del 2013 vaig tenir l'oportunitat de realitzar una estada en el laboratori del Prof. Jason Lieb a la Universitat de Carolina del Nord, EE.UU. En dita col·laboració es volien estudiar els patrons de digestió de la nucleasa micrococal, d'una forma similar en com havíem presentat en els nostres treballs anteriors. Per fer-ho, es van comparar experiments amb diferents nivells de digestió en C.Elegans i es van estudiar els patrons observats. Alguns dels resultats presentats inclouen patrons de tall similars als observats per nosaltres prèviament en llevat, així com la seva relació amb seqüències amb una caracterització física anormal. També es van veure com diferents regions es comporten de forma diferent a la digestió, podent trobar nucleosomes que tenen una tendència a ser sempre més resistents, menys resistents o que mostren una sensibilitat dinàmica a diferents nivells de digestió.

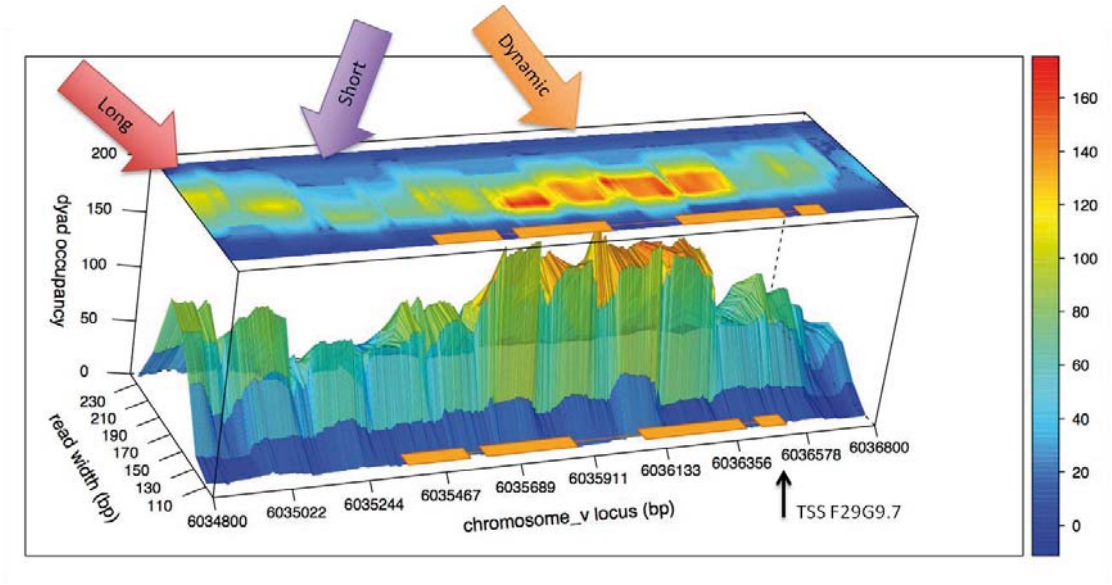


Figura 12. Plot de cobertura en 3D al voltant del gen F29G9. La profunditat (eix Y), mostra la concentració de molècules de diferents mides (més o menys digerides). Les fletxes assenyalen nucleosomes individuals amb diferents respostes als canvis de agressivitat en la digestió. En la part superior de la caixa, es mostra el mateix en 2D, seguint un patró similar a un gel d'agarosa simulat.

4. Discussió i Conclusions

4.1. Discussió

En la primer bloc de publicacions hem parlat de l'impacte de les propietats físiques de l'ADN sobre diferents resultats experimentals. La combinació de diferents factors *in vivo* (remodeladors de cromatina, factors de transcripció, modificacions epigenètiques, etc.) amb aquestes propietats físiques assenyalen llocs crítics per la regulació genètica. Més enllà del conegut patró de digestió al voltant dels inicis de transcripció, en els nostres treballs hem vist com aquestes descripcions físiques permeten modelar el comportament enzimàtic de la MNasa, la metilació de l'ADN o detectar noves regions amb activitat transcripcional en humans.

El potencial d'aquestes propietats físiques presenta, doncs , una nova dimensió en la comprensió de mecanismes moleculars de baix nivell, la predicció del posicionament dels nucleosomes o fins i tot la seva utilització per la predicció de l'eficiència de nous fàrmacs (Sciabola *et al.*, 2013). Els treballs aquí presentats serviran de base pel desenvolupament d'aquests nous models, els quals necessiten noves aproximacions experimentals per entendre l'efecte concret de la metilació de l'ADN sobre el posicionament de nucleosomes o estructures de nucleosomes no canòniques.

Tot i que el laboratori del EBL té una aparició recent, alguns dels articles presentats demostren que comença a assolir una maduresa que es farà encara més patent en els propers anys. Més enllà de la simple validació experimental dels resultats teòrics proposats en el laboratori del MMB, presentem aquí diferents col·laboracions tan internes com externes que combinen l'excel·lència tan en les aproximacions experimental com en l'anàlisi bioinformàtic. Durant els anys que comprenen aquesta tesis, hem seqüenciat i analitzat més de 60 mapes de nucleosomes. Això ens ha permès entendre les fonts de variabilitat tant tècnica com biològica en dits mapes i estudiar la dinàmica de la cromatina al llarg del cicle cel·lular, presentant noves i rellevants contribucions cap a una millor comprensió de la cromatina i la seva regulació.

La recerca feta al laboratori s'ha complementat amb diverses col·laboracions amb grups de renom internacional, permetent l'aplicació de les tècniques desenvolupades en el nostre grup sobre altres organismes o processos biològics.

Evidentment, al llarg d'aquesta recerca s'han obert nous interrogants sobre aspectes més complexos de la cromatina, com estructures no canòniques, el paper de les variants d'histones, diferents modificacions de l'ADN o llocs amb una caracterització especial, com els telòmers o els centròmers. Alguns dels resultats futurs en aquest camp segurament puguin tenir un impacte directe sobre fenotips clínics o el disseny de noves solucions biomèdiques, reduint així la bretxa entre la recerca bàsica i aplicada.

Com a últim punt, una part important dels resultats presentats aquí corresponen a noves eines i mètodes bioinformàtics que permeten l'anàlisi o la integració d'experiments de diferent naturalesa. En general, observem una tendència a estudiar cada cop més els detalls en comptes de la visió general d'un determinat problema. Així doncs les eines bioinformàtiques s'han de concentrar en identificar quines parts del senyals observats corresponen a una variació real i quines és poden deure a artefactes diversos. La simplificació dels processos de pre-anàlisis i la interconnexió amb bases de dades externes també és un repte que hem intentat adreçar. El creixement del volum i detall de les dades a analitzar i l'evolució dels mitjans computacionals permeten nous anàlisis integratius fins ara inimaginables però presenten limitacions tècniques que requereixen la confecció de programari especialitzat i conscient de l'arquitectura. Esperem que les eines aquí presentades contribueixin en alguna mesura en assolir aquests objectius.

4.2. Conclusions generals

- Els nucleosomes són entitats altament dinàmiques i l'estudi del seu posicionament requereix tan d'un disseny experimental sistemàtic com un anàlisis bioinformàtic precís.
- Les propietats físiques de l'ADN sou un actor essencial en diferents processos tan *in vivo* com *in vitro*. Hem de tenir en compte aquesta nova dimensió d'anàlisis en la mateixa mesura que utilitzem altres factors experimentals per explicar diferents processos biològics.
- La ràpida evolució de les tecnologies d'alt rendiment en genòmica i epigenòmica requereixen eines de processament eficients i fiables que puguin manipular grans volums de dades sense intervenció humana en els anàlisis preliminars.
- La interacció entre nucleosomes, modificacions de l'ADN, les propietats físiques dels àcids nucleics i les estructures anormals de la cromatina son un tema d'estudi clau per entendre diferents processos biològics bàsics, així com per avançar cap a la definició, i per tant el tractament, de diferents malalties

REFERENCES

Adams,D. *et al.* (2012) BLUEPRINT to decode the epigenetic signature written in blood. *Nat. Biotechnol.*, **30**, 224–6.

Aird,D. *et al.* (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.*, **12**, R18.

Alberts *et al.* (2004) Essential Cell Biology 2nd ed. Garland Science.

Allan,J. *et al.* (2012) Micrococcal nuclease does not substantially bias nucleosome mapping. *J. Mol. Biol.*, **417**, 152–64.

Bähler,J. (2005) Cell-cycle control of gene expression in budding and fission yeast. *Annu. Rev. Genet.*, **39**, 69–94.

Balasubramanian,S. *et al.* (2009) DNA sequence-directed organization of chromatin: structure-based computational analysis of nucleosome-binding sequences. *Biophys. J.*, **96**, 2245–60.

Battistini,F. *et al.* (2010) Structural mechanics of DNA wrapping in the nucleosome. *J. Mol. Biol.*, **396**, 264–79.

Baù,D. *et al.* (2011) The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. *Nat. Struct. Mol. Biol.*, **18**, 107–14.

Becker,J. *et al.* (2013) NucleoFinder: a statistical approach for the detection of nucleosome positions. *Bioinformatics*, **29**, 711–6.

Bell,O. *et al.* (2011) Determinants and dynamics of genome accessibility. *Nat. Rev. Genet.*, **12**, 554–564.

Benjamini,Y. and Speed,T.P. (2012) Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.*, **40**, e72.

Bergman,Y. and Cedar,H. (2013) DNA methylation dynamics in health and disease. *Nat. Struct. Mol. Biol.*, **20**, 274–81.

Bernstein,B.E. *et al.* (2010) The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.*, **28**, 1045–8.

Bird,A. (2007) Perceptions of epigenetics. *Nature*, **447**, 396–8.

Brogaard,K. *et al.* (2012) A map of nucleosome positions in yeast at base-pair resolution. *Nature*.

Cairns,B.R. (2009) The logic of chromatin architecture and remodelling at promoters. *Nature*, **461**, 193–8.

Cao,H. *et al.* (1998) TGGA repeats impair nucleosome formation. *J. Mol. Biol.*, **281**, 253–60.

Cedar,H. and Bergman,Y. (2009) Linking DNA methylation and histone modification: patterns and paradigms. *Nat. Rev. Genet.*, **10**, 295–304.

Cedar,H. and Bergman,Y. (2012) Programming of DNA methylation patterns. *Annu. Rev. Biochem.*, **81**, 97–117.

Chen,K. *et al.* (2013) DANPOS: dynamic analysis of nucleosome position and occupancy by sequencing. *Genome Res.*, **23**, 341–51.

Chen,T. and Dent,S.Y.R. (2014) Chromatin modifiers and remodellers: regulators of cellular differentiation. *Nat. Rev. Genet.*, **15**, 93–106.

Chen,X. *et al.* (2010) A dynamic Bayesian network for identifying protein-binding footprints from single molecule-based sequencing data. *Bioinformatics*, **26**, i334–i342.

Chodavarapu,R.K. *et al.* (2010) Relationship between nucleosome positioning and DNA methylation. *Nature*, **466**, 388–392.

Chung,H.-R. *et al.* (2010a) The effect of micrococcal nuclease digestion on nucleosome positioning data. *PLoS One*, **5**, e15754.

Chung,H.-R. *et al.* (2010b) The effect of micrococcal nuclease digestion on nucleosome positioning data. *PLoS One*, **5**, e15754.

Chung,H.-R. and Vingron,M. (2009) Sequence-dependent nucleosome positioning. *J. Mol. Biol.*, **386**, 1411–22.

Collings,C.K. *et al.* (2013) Effects of DNA methylation on nucleosome stability. *Nucleic Acids Res.*, **41**, 2918–31.

Crick,F. (1970) Central dogma of molecular biology. *Nature*, **227**, 561–3.

Cui,F. and Zhurkin,V.B. (2010) Structure-based analysis of DNA sequence patterns guiding nucleosome positioning in vitro. *J. Biomol. Struct. Dyn.*, **27**, 821–41.

Dabney,J. and Meyer,M. (2012) Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *Biotechniques*, **52**, 87–94.

Davey,C.A. *et al.* (2002) Solvent Mediated Interactions in the Structure of the Nucleosome Core Particle at 1.9Å Resolution†. *J. Mol. Biol.*, **319**, 1097–1113.

Deniz,Ö. *et al.* (2011) Physical properties of naked DNA influence nucleosome positioning and correlate with transcription start and termination sites in yeast. *BMC Genomics*, **12**, 489.

Dickerson,R.E. *et al.* (1989) Definitions and nomenclature of nucleic acid structure parameters. *EMBO J.*, **8**, 1–4.

Dunham,I. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Ecker,J.R. *et al.* (2012) Genomics: ENCODE explained. *Nature*, **489**, 52–5.

Egger,G. *et al.* (2004) Epigenetics in human disease and prospects for epigenetic therapy. *Nature*, **429**, 457–63.

Ernst,J. and Kellis,M. (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.*, **28**, 817–825.

Esteller,M. and Herman,J.G. (2002) Cancer as an epigenetic disease: DNA methylation and chromatin alterations in human tumours. *J. Pathol.*, **196**, 1–7.

Field,Y. *et al.* (2008) Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS Comput. Biol.*, **4**, e1000216.

Finch,J.T. *et al.* (1977) Structure of nucleosome core particles of chromatin. *Nature*, **269**, 29–36.

Flaus,A. *et al.* (1996) Mapping nucleosome position at single base-pair resolution by using site-directed hydroxyl radicals. *Proc. Natl. Acad. Sci.*, **93**, 1370–1375.

Flick,J.T. *et al.* (1986) Micrococcal nuclease as a DNA structural probe: Its recognition sequences, their genomic distribution and correlation with DNA structure determinants. *J. Mol. Biol.*, **190**, 619–633.

Flores,O. *et al.* (2009) New multiplatform computer program for numerical identification of microorganisms. *J. Clin. Microbiol.*, **47**, 4133–5.

Gabdank,I. *et al.* (2010) FineStr: a web server for single-base-resolution nucleosome positioning. *Bioinformatics*, **26**, 845–6.

Gentleman,R.C. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Di Gesù,V. *et al.* (2009) A multi-layer method to study genome-scale positions of nucleosomes. *Genomics*, **93**, 140–5.

Goñi,J.R. *et al.* (2007) Determining promoter location based on DNA structure first-principles calculations. *Genome Biol.*, **8**, R263.

Goñi,J.R. *et al.* (2008) DNAlive: a tool for the physical analysis of DNA at the genomic scale. *Bioinformatics*, **24**, 1731–2.

Grewal,S.I.S. and Moazed,D. (2003) Heterochromatin and epigenetic control of gene expression. *Science*, **301**, 798–802.

Guan,D. *et al.* (2013) Switching cell fate, ncRNAs coming to play. *Cell Death Dis.*, **4**, e464.

Gupta,S. *et al.* (2008) Predicting human nucleosome occupancy from primary sequence. *PLoS Comput. Biol.*, **4**, e1000134.

Hannon,G.J. (2002) RNA interference. *Nature*, **418**, 244–51.

Hogan,G.J. *et al.* (2006) Cell cycle-specified fluctuation of nucleosome occupancy at gene promoters. *PLoS Genet.*, **2**, e158.

Hörz,W. and Altenburger,W. (1981) Sequence specific cleavage of DNA by micrococcal nuclease. *Nucleic Acids Res.*, **9**, 2643–58.

Huff,J.T. and Zilberman,D. (2014) Dnmt1-Independent CG Methylation Contributes to Nucleosome Positioning in Diverse Eukaryotes. *Cell*, **156**, 1286–97.

Ioshikhes,I. *et al.* (2011) Variety of genomic DNA patterns for nucleosome positioning. *Genome Res.*, gr.116228.110–.

Ioshikhes,I.P. *et al.* (2006) Nucleosome positions predicted through comparative genomics. *Nat. Genet.*, **38**, 1210–5.

Jenuwein,T. and Allis,C.D. (2001) Translating the histone code. *Science*, **293**, 1074–80.

Jiang,C. and Pugh,B.F. (2009) Nucleosome positioning and gene regulation: advances through genomics. *Nat. Rev. Genet.*, **10**, 161–72.

Jones,P.A. (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.*, **13**, 484–92.

Jones,P.A. and Liang,G. (2009) Rethinking how DNA methylation patterns are maintained. *Nat. Rev. Genet.*, **10**, 805–11.

Kamakaka,R.T. and Biggins,S. (2005) Histone variants: deviants? *Genes Dev.*, **19**, 295–310.

Kaplan,N. *et al.* (2010) Nucleosome sequence preferences influence in vivo nucleosome organization. *Nat. Struct. Mol. Biol.*, **17**, 918–920.

Kaplan,N. *et al.* (2009) The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, **458**, 362–6.

Kelly,T.K. *et al.* (2012) Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res.*, **22**, 2497–506.

Kornberg,R.D. and Lorch,Y. (1999) Twenty-five years of the nucleosome, review fundamental particle of the eukaryote chromosome. *Cell*, **98**, 285–294.

Kornberg,R.D. and Stryer,L. (1988) Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism. *Nucleic Acids Res.*, **16**, 6677–90.

Kuan,P.F. *et al.* (2009) A non-homogeneous hidden-state model on first order differences for automatic detection of nucleosome positions. *Stat. Appl. Genet. Mol. Biol.*, **8**, Article29.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Lankas,F. *et al.* (2003) DNA basepair step deformability inferred from molecular dynamics simulations. *Biophys. J.*, **85**, 2872–83.

Lavery,R. *et al.* (2010a) 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**, 299–313.

Lavery,R. *et al.* (2010b) 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**, 299–313.

Law,J.A. and Jacobsen,S.E. (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.*, **11**, 204–20.

Lee,W. *et al.* (2007) A high-resolution atlas of nucleosome occupancy in yeast. *Nat. Genet.*, **39**, 1235–44.

Levitt,M. (1983) Computer Simulation of DNA Double-helix Dynamics. *Cold Spring Harb. Symp. Quant. Biol.*, **47**, 251–262.

Li,Z. *et al.* (2011) The nucleosome map of the mammalian liver. *Nat. Struct. Mol. Biol.*, **18**, 742–746.

Liu,L. *et al.* (2012) Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.*, **2012**, 251364.

Loman,N.J. *et al.* (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.*, **30**, 434–9.

Lowary,P.T. and Widom,J. (1998) New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *J. Mol. Biol.*, **276**, 19–42.

Lu,X.-J. and Olson,W. (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.*, **31**, 5108–5121.

Lubliner,S. and Segal,E. (2009) Modeling interactions between adjacent nucleosomes improves genome-wide predictions of nucleosome occupancy. *Bioinformatics*, **25**, i348–55.

Luger,K. *et al.* (1997) Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature*, **389**, 251–60.

Lutter,L.C. (1989) Digestion of nucleosomes with deoxyribonucleases I and II. *Methods Enzymol.*, **170**, 264–9.

Marco-Sola,S. *et al.* (2012) The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat. Methods*, **9**, 1185–8.

Mavrich,T.N., Ioshikhes,I.P., *et al.* (2008) A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res.*, **18**, 1073–83.

Mavrich,T.N., Jiang,C., *et al.* (2008) Nucleosome organization in the Drosophila genome. *Nature*, **453**, 358–62.

Milani,P. *et al.* (2009) Nucleosome positioning by genomic excluding-energy barriers. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 22257–62.

Möbius,W. *et al.* (2013) Toward a unified physical model of nucleosome patterns flanking transcription start sites. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, 5719–24.

Morozov,A. V *et al.* (2009) Using DNA mechanics to predict in vitro nucleosome positions and formation energies. *Nucleic Acids Res.*, **37**, 4707–22.

Nellore,A. *et al.* (2012) NSeq: a multithreaded Java application for finding positioned nucleosomes from sequencing data. *Front. Genet.*, **3**, 320.

Nikitina,T. *et al.* (2013) Combined micrococcal nuclease and exonuclease III digestion reveals precise positions of the nucleosome core/linker junctions: implications for high-resolution nucleosome mapping. *J. Mol. Biol.*, **425**, 1946–60.

Noy,A. *et al.* (2005) Structure, recognition properties, and flexibility of the DNA.RNA hybrid. *J. Am. Chem. Soc.*, **127**, 4910–20.

Olson,W.K. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci.*, **95**, 11163–11168.

Olson,W.K. and Zhurkin,V.B. (2011) Working the kinks out of nucleosomal DNA. *Curr. Opin. Struct. Biol.*, **21**, 348–57.

Orozco,M. *et al.* (2008) Recent advances in the study of nucleic acid flexibility by molecular dynamics. *Curr. Opin. Struct. Biol.*, **18**, 185–93.

Pabo,C.O. and Sauer,R.T. (1984) Protein-DNA recognition. *Annu. Rev. Biochem.*, **53**, 293–321.

Park,P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, **10**, 669–80.

Parmar,J.J. *et al.* (2013) Nucleosome positioning and kinetics near transcription-start-site barriers are controlled by interplay between active remodeling and DNA sequence. *Nucleic Acids Res.*

Patel,R.K. and Jain,M. (2012) NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data. *PLoS One*, **7**, e30619.

Peckham,H.E. *et al.* (2007) Nucleosome positioning signals in genomic DNA. *Genome Res.*, **17**, 1170–7.

Pérez,A., Luque,F.J., *et al.* (2007) Dynamics of B-DNA on the microsecond time scale. *J. Am. Chem. Soc.*, **129**, 14739–45.

Pérez,A., Marchán,I., *et al.* (2007) Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J.*, **92**, 3817–29.

Pérez,A. *et al.* (2008) Towards a molecular dynamics consensus view of B-DNA flexibility. *Nucleic Acids Res.*, **36**, 2379–94.

Polishko,A. *et al.* (2012) NORMAL: accurate nucleosome positioning using a modified Gaussian mixture model. *Bioinformatics*, **28**, i242–9.

Portela,A. *et al.* (2013) DNA methylation determines nucleosome occupancy in the 5'-CpG islands of tumor suppressor genes. *Oncogene*.

Portela,A. and Esteller,M. (2010) Epigenetic modifications and human disease. *Nat. Biotechnol.*, **28**, 1057–1068.

Richmond,T.J. *et al.* (1984) Structure of the nucleosome core particle at 7 Å resolution. *Nature*, **311**, 532–537.

Richmond,T.J. and Davey,C.A. (2003) The structure of DNA in the nucleosome core. *Nature*, **423**, 145–50.

Rivera,C.M. and Ren,B. (2013) Mapping human epigenomes. *Cell*, **155**, 39–55.

Sanger,F. and Coulson,A.R. (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.*, **94**, 441–8.

De Santis,P. and Scipioni,A. (2013) Sequence-dependent collective properties of DNAs and their role in biological systems. *Phys. Life Rev.*, **10**, 41–67.

Satchwell,S.C. *et al.* (1986) Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.*, **191**, 659–675.

Schones,D.E. *et al.* (2008) Dynamic regulation of nucleosome positioning in the human genome. *Cell*, **132**, 887–98.

Schöpflin,R. *et al.* (2013) Modeling nucleosome position distributions from experimental nucleosome positioning maps. *Bioinformatics*.

Sciabola,S. *et al.* (2013) Improved nucleic acid descriptors for siRNA efficacy prediction. *Nucleic Acids Res.*, **41**, 1383–94.

Segal,E. *et al.* (2006) A genomic code for nucleosome positioning. *Nature*, **442**, 772–8.

Segal,E. and Widom,J. (2009) What controls nucleosome positions? *Trends Genet.*, **25**, 335–43.

Shivaswamy,S. *et al.* (2008) Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biol.*, **6**, e65.

Skene,P.J. and Henikoff,S. (2013) Histone variants in pluripotency and disease. *Development*, **140**, 2513–24.

Smith,Z.D. and Meissner,A. (2013) DNA methylation: roles in mammalian development. *Nat. Rev. Genet.*, **14**, 204–20.

Solomon,M.J. *et al.* (1988) Mapping proteinDNA interactions in vivo with formaldehyde: Evidence that histone H4 is retained on a highly transcribed gene. *Cell*, **53**, 937–947.

Spies,N. *et al.* (2009) Biased chromatin signatures around polyadenylation sites and exons. *Mol. Cell*, **36**, 245–54.

Stein,A. *et al.* (2010) Are nucleosome positions in vivo primarily determined by histone-DNA sequence preferences? *Nucleic Acids Res.*, **38**, 709–19.

Struhl,K. and Segal,E. (2013) Determinants of nucleosome positioning. *Nat. Struct. Mol. Biol.*, **20**, 267–73.

Sun,W. *et al.* (2009) Dissecting nucleosome free regions by a segmental semi-Markov model. *PLoS One*, **4**, e4721.

Teif,V.B. and Rippe,K. (2009) Predicting nucleosome positions on the DNA: combining intrinsic sequence preferences and remodeler activities. *Nucleic Acids Res.*, **37**, 5641–55.

Teng,Y. *et al.* (2001) The mapping of nucleosomes and regulatory protein binding sites at the Saccharomyces cerevisiae MFA2 gene: a high resolution approach. *Nucleic Acids Res.*, **29**, E64–4.

Thåström,A. *et al.* (1999) Sequence motifs and free energies of selected natural and non-natural nucleosome positioning DNA sequences. *J. Mol. Biol.*, **288**, 213–29.

Tidor,B. *et al.* (1983) Dynamics of DNA oligomers. *J. Biomol. Struct. Dyn.*, **1**, 231–52.

Tilgner,H. *et al.* (2009) Nucleosome positioning as a determinant of exon recognition. *Nat. Struct. Mol. Biol.*, **16**, 996–1001.

Tillo,D. and Hughes,T.R. (2009) G+C content dominates intrinsic nucleosome occupancy. *BMC Bioinformatics*, **10**, 442.

Tolstorukov,M.Y. *et al.* (2007) A novel roll-and-slide mechanism of DNA folding in chromatin: implications for nucleosome positioning. *J. Mol. Biol.*, **371**, 725–38.

Trifonov,E.N. (2010a) Base pair stacking in nucleosome DNA and bendability sequence pattern. *J. Theor. Biol.*, **263**, 337–9.

Trifonov,E.N. (2010b) Nucleosome positioning by sequence, state of the art and apparent finale. *J. Biomol. Struct. Dyn.*, **27**, 741–6.

Tsankov,A. *et al.* (2011) Evolutionary divergence of intrinsic and trans-regulated nucleosome positioning sequences reveals plastic rules for chromatin organization. *Genome Res.*, **21**, gr.122267.111–.

Valouev,A. *et al.* (2008) A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Res.*, **18**, 1051–63.

Varley,K.E. *et al.* (2013) Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res.*, **23**, 555–67.

Vasudevan,D. *et al.* (2010) Crystal Structures of Nucleosome Core Particles Containing the “601” Strong Positioning Sequence. *J. Mol. Biol.*, **403**, 1–10.

Waddington,C.H. (1959) Canalization of Development and Genetic Assimilation of Acquired Characters. *Nature*, **183**, 1654–1655.

Watson,J.D. and Crick,F.H. (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, **171**, 737–8.

White,O. *et al.* (1993) A quality control algorithm for DNA sequencing projects. *Nucleic Acids Res.*, **21**, 3829–3838.

Wu,R. and Li,H. (2010) Positioned and G/C-capped poly(dA:dT) tracts associate with the centers of nucleosome-free regions in yeast promoters. *Genome Res.*, **20**, 473–84.

Xi,L. *et al.* (2010) Predicting nucleosome positioning using a duration Hidden Markov Model. *BMC Bioinformatics*, **11**, 346.

Xu,F. and Olson,W.K. (2010) DNA architecture, deformability, and nucleosome positioning. *J. Biomol. Struct. Dyn.*, **27**, 725–39.

Yamasaki,S. *et al.* (2009) A generalized conformational energy function of DNA derived from molecular dynamics simulations. *Nucleic Acids Res.*, **37**, e135.

Yassour,M. *et al.* (2008) Nucleosome positioning from tiling microarray data. *Bioinformatics*, **24**, i139–46.

Yen,K. *et al.* (2012) Genome-wide nucleosome specificity and directionality of chromatin remodelers. *Cell*, **149**, 1461–73.

Yuan,G.-C. *et al.* (2005) Genome-scale identification of nucleosome positions in S. cerevisiae. *Science*, **309**, 626–30.

Zaugg,J.B. and Luscombe,N.M. (2011) A genomic model of condition-specific nucleosome behaviour explains transcriptional activity in yeast. *Genome Res.*, **22**, gr.124099.111–.

Zentner,G.E. and Henikoff,S. (2013) Regulation of nucleosome dynamics by histone modifications. *Nat. Struct. Mol. Biol.*, **20**, 259–66.

Zhang,X. *et al.* (2012) Probabilistic inference for nucleosome positioning with MNase-based or sonicated short-read data. *PLoS One*, **7**, e32095.

Zhang,Y. *et al.* (2009) Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. *Nat. Struct. Mol. Biol.*, **16**, 847–52.

Zhang,Y. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.

Zlatanova,J. *et al.* (2009) The nucleosome family: dynamic and growing. *Structure*, **17**, 160–71.

ANNEX 1: BREU INTRODUCCIÓ A LA BIOLOGIA MOLECULAR PER A NO BIÒLEGS

Si bé la present tesis doctoral s'emmarca dins el camp de les ciències de la vida i la salut, la meva formació prèvia poc té a veure amb aquest terreny. Com a enginyer informàtic, els meus coneixements de biomedicina o biologia en general al començar el doctorat eren pràcticament nuls. Per tant, aquests coneixements bàsics també els he adquirit durant la tesi i m'agradaria incloure'ls aquí, si bé la seva generalitat no justifica l'esforç de traducció ni la seva inclusió en la memòria principal. Penso que aquestes nocions poden guiar tan a futurs investigadors que, com jo, s'introdueixin a l'estudi bioinformàtic de la cromatina provinent d'una formació tècnica com a qui tingui un interès personal en aquesta tesi i no tingui coneixements previs de biologia.

1. ADN: L'estructura fundamental de la vida

Definir el terme "vida" és una tasca complexa i no exempta de polèmica. En tot cas, una bona definició hauria d'incloure dos elements essencials: el caràcter dinàmic i perpetuable de la mateixa juntament amb un cert grau d'interacció i adaptació amb el medi que l'envolta. No tot el que es mou ni tot el que interacciona és viu, però lluny d'entrar en un debat semiològic el que es pretén posar en relleu és que la vida es fonamenta en un "moviment autosostenible".

Aquest moviment no sempre té unes causes clares. És veritat que moltes de les reaccions que donen lloc a la vida troben el seu origen en un manteniment predictible de l'equilibri físic o químic entre àtoms o molècules, però també hi ha fenòmens aparentment aleatoris que afecten un sistema viu. Tots hem sentit a parlar de la teoria de l'evolució de Charles Darwin, la qual apunta que els trets presents en individus amb unes capacitats més grans de supervivència acaben transmetent-se a l'espècie a traves d'una selecció natural. Aquesta transmissió intergeneracional de trets específics troba les seves bases biològiques en els gens.

Conceptualment un **gen** es pot entendre com una unitat bàsica d'informació biològica, implementada físicament sobre una molècula amb unes propietats químiques concretes que li donen el nom tècnic d'**àcid desoxiribonucleic**, més conegut com a **ADN**.

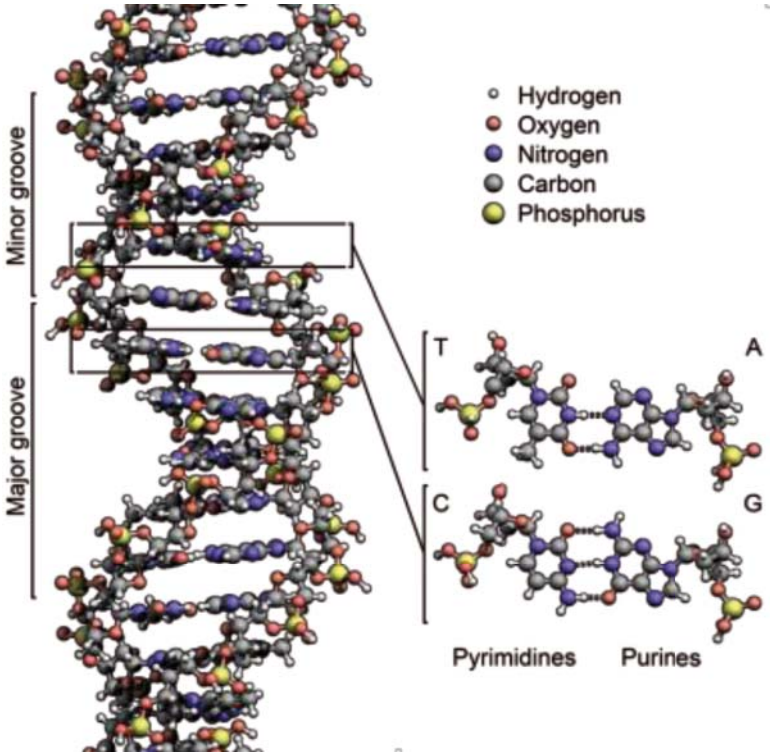


Fig. Annx. 1. Estructura bàsica de l'ADN

Com es pot observar a la Fig. Annx. 1, l'estructura fonamental de l'ADN es basa una combinació d'àcids nucleics o nucleòtids. Aquests àcids nucleics son petites molècules amb una gran capacitat per formar polímers –cadenes– estables deguda la seva estructura atòmica. L'estructura d'un nucleòtid es defineix per un monosacàrid –un tipus bàsic de sucre–, una base nitrogenada i un grup fosfat que permet l'enllaç entre bases. A efectes pràctics, centrant-nos només en l'ADN, el que diferencia un nucleòtid d'un altre és la seva base nitrogenada, que pot ser una entre **adenina** (A), **citosina** (C), **guanina** (G) o **timina** (T). La combinació seqüencial d'aquestes 4 lletres (ACGT) serà el que posteriorment definirà el **genoma** –el conjunt de gens– d'un organisme.

L'estructura de l'ADN es troba caracteritzada per una doble hèlix conformada per dues cadenes de nucleòtids. Quan parlem de la *seqüència del genoma* d'un organisme, ens referim a la seqüència de bases que trobem en una sola d'aquestes cadenes, doncs la cadena oposada sempre es complementària a la primera. Aquesta complementarietat es dona perquè, en condicions normals, una adenina (A) sempre forma l'enllaç complementari amb una timina (T), i una citosina (C) sempre s'unirà a una guanina (G), almenys en l'ADN. Això es deu a l'estructura bioquímica de les bases, on una purina (A, G) s'enllaça preferentment amb una pirimidina (C, T) degut al nombre d'enllaços d'hidrogen que poden formar (A-T en formen 2 i C-G en formen 3).

una purina (A, G) s'enllaça preferentment amb una pirimidina (C, T) degut al nombre d'enllaços d'hidrogen que poden formar (A-T en formen 2 i C-G en formen 3).

Es fa difícil però pensar com a partir d'aquesta seqüència de nucleòtids es pugui donar lloc a una cosa tan complexa com és un ésser viu. Podem fer un símil amb l'electrònica present en un ordinador: quan mirem una fotografia en pantalla, nosaltres no veiem ni podem entendre el procés pel qual els diferents electrons s'organitzen en els transistors per formar bits i aquests bits donen lloc als colors que veiem per pantalla. Quin és, doncs, el procés bàsic pel qual una modificació en aquesta seqüència de nucleòtids definirà el color dels nostres ulls o el risc de patir una determinada malaltia?

Si bé la seqüència genòmica serveix com a pauta per definir el que acabarà materialitzant-se en els diferents processos metabòlics, els gens, entesos al nivell de l'ADN, no tenen capacitat funcional. El pas de la seqüència de nucleòtids a una unitat que té una determinada funció s'explica a través del que s'anomena el **dogma central de la biologia molecular** (Fig. Annx. 2).

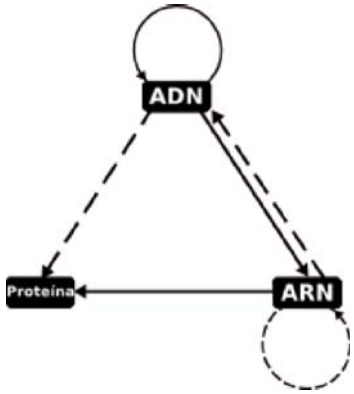


Fig. Annx. 2. Dogma central de la biologia molecular, versió revisada al 1970.

Fins el 1970 una visió reduccionista d'aquest dogma definia que la diferent maquinària cel·lular s'encarregava de traduir l'ADN en ARN i al seu temps l'ARN era transcrit en proteïnes. **ARN** és l'acrònim d'**àcid ribonucleic**, un altre tipus de material genètic format per àcids nucleics amb unes propietats lleugerament diferents a les de l'ADN. Entre aquestes propietats diferencials cal destacar que en el procés de traducció de ADN a ARN, la timina (T) es converteix en **uracil** (U, una altra pirimidina), pel que el codi genètic del ARN està format per les lletres ACGU. Un altre diferencia significativa rau en que normalment les seqüències d'ARN son de cadena simple (no acostumen a formar una doble hèlix). Tot i així, com hem dit, al 1970 es va descobrir que el paper del ARN anava molt

més enllà de ser un simple missatger entre l'ADN i la maquinaria cel·lular encarregada de la síntesis de proteïnes. L'ARN pot influir en la replicació de l'ADN, pot interferir amb altres molècules d'ARN i inclús pot autoreplicar-se en alguns organismes que no necessiten ADN, com els virus. Cal també esmenar que en el pas d'ADN a ARN no és produeix una còpia idèntica. En l'ADN d'un gen donat trobem diferents regions dins la seva seqüència que contenen informació sobre la codificació de proteïnes, anomenades **exons**, i altres regions no traduïdes, anomenades **introns** (Fig. Annx. 3).

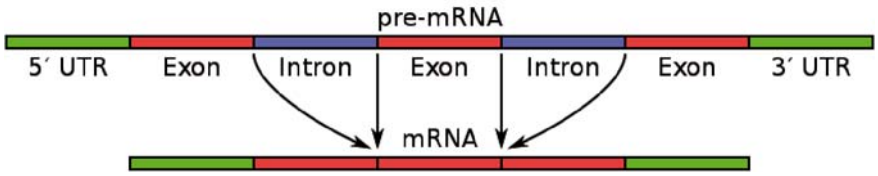


Fig. Annx. 3. Traducció d'ADN a ARN pel procés de *splicing*

En el procés de traducció d'ADN a ARN, anomenat ***splicing***, la seqüència dels diferents exons es tradueix i es concatena, mentre que els introns es descarten. Relacionat en aquest punt, val la pena mencionar que no tots els exons presents en la seqüència d'ADN d'un gen s'han de traduir a la vegada. Això dona lloc a un fenomen anomenat ***splicing* alternatiu**, que permet que puguin existir variants d'un mateix gen –isoformes– amb diferents funcions.

No entrarem en els mecanismes concrets de transcripció i síntesi de proteïnes a partir del ADN/ARN, però val la pena comentar que la maquinària que s'encarrega de traduir o replicar l'ADN i l'ARN s'anomena **polimerasa** i que el mecanisme que agrupa els diferents aminoàcids a partir del patró que li proporciona l'ARN és diu **ribosoma**.

L'últim punt que cal comentar per introduir les estructures bàsiques sobre les que es fonamenta la vida són les proteïnes.

En general, anomenem **proteïna** a una cadena d'aminoàcids units mitjançant enllaços peptídics. Un **aminoàcid** és un compost orgànic format per un grup amino, un grup carboxil i una cadena lateral que li dona nom. A pesar de que aquesta denominació permet moltes combinacions, en humà només en trobem 22 i cadascun d'ells té un o més triplets d'ARN –**codons**– que el codifiquen (Fig. Annx. 4). Per exemple, si en la nostra cadena d'ARN trobem la seqüència de nucleòtids **GCA** aquella seqüència codificarà per un aminoàcid anomenat alanina (*Ala*, en la notació abreviada), mentre que **AAA** codificaria per una lisina (*Lys*) (Fig. Annx. 4). Un tipus especial de codons son els que marquen

el inici o el final de transcripció, és a dir, que la polimerasa detectarà aquells patrons i iniciarà o acabarà la transcripció a partir d'aquell punt. En humà, el codó **AUG** marca el inici de transcripció al mateix temps que codifica per l'aminoàcid metionina (*Met*) i els finals de transcripció poden venir donats pels codons **UAA**, **UAG** o **UGA**.

A) **B)**

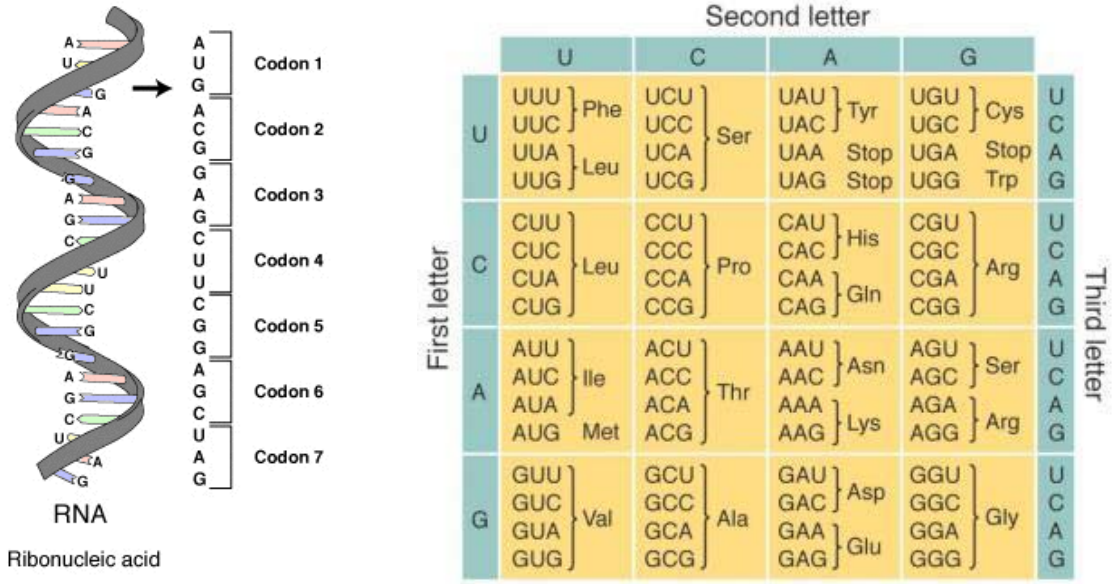


Fig. Annx. 4. A) Representació d'una molècula de ARN i la seva organització en codons. B) Taula d'equivalències entre els diferents codons i els codis dels aminoàcid associats.

La cadena d'aminoàcids d'una proteïna s'anomena estructura primària, doncs és només un primer nivell de detall de la forma final de la proteïna. Els diferents àtoms dels grups amino i carbonil dels aminoàcids interaccionen i s'enllacen entre ells formant un plegament local que defineix el que s'anomena l'estructura secundaria de la proteïna. Les estructures secundàries més comuns son les hèlix alfa i les fulles beta, anomenades així per la posició del carboni que forma l'enllaç dins la molècula. Posteriorment trobem un plegament a nivell global de la proteïna on les regions polars (amb diferent electronegativitat) s'organitzen cap a l'exterior per interactuar amb les molècules d'aigua que l'envolten i les regions apolars (amb igual electronegativitat) s'organitzen cap a l'interior de la proteïna. Aquesta conformació defineix el que s'anomena estructura terciària de la proteïna. Finalment, també podem trobar proteïnes formades per més d'una cadena d'aminoàcids interactuant entre elles. L'estructura composta entre diferents cadenes s'anomena estructura quaternària.

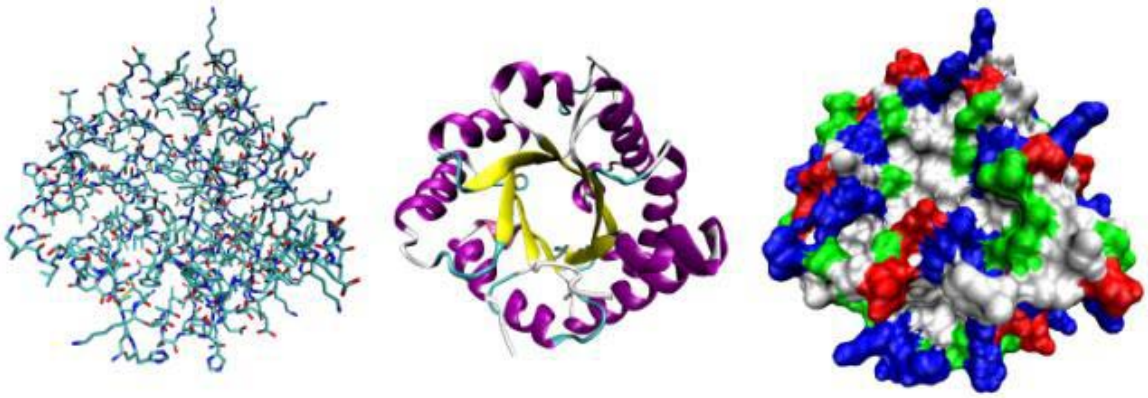


Fig. Annx. 5. Diferents representacions de la estructura tridimensional de la proteïna *triosafosfat isomerasa.* A la dreta les unions entre els diferents àtoms es representen com a palets de colors. Al centre les hèlixs alfa i les lamines beta apareixen en una representació esquemàtica en porpra i groc respectivament. A l'esquerra es mostra la superfície accessible al solvent, amb els residus àcids en vermell, els bàsics en blau, els polars en verd i els apolars en blanc.

El conjunt de proteïnes d'un organisme s'anomena **proteoma**. De proteïnes n'hi ha de molts tipus, amb moltes formes, mides i funcions diferents. Hi ha proteïnes estructurals, com el col·lagen, que formen els nostres ossos i teixits. Hi ha proteïnes contràctils, com l'actina, que permet la contracció de les parets cel·lulars i per extensió dels nostres músculs. Hi ha proteïnes enzimàtiques, com la triosafosfat isomerasa (Fig. Annx. 5), involucrada en el procés de transformació de sucres en energia dins les nostres cèl·lules. Hi ha proteïnes que sintetitzen proteïnes, com les que formen part del ribosoma. Hi ha proteïnes involucrades en qualsevol procés que passa dins un ésser viu.

Fins aquí hem vist les bases bioquímiques de les estructures fonamentals que donen lloc a la vida i també com la seqüència genòmica serveix de motlle sobre el qual es generen les proteïnes que acabaran influint en cadascun dels nostres processos metabòlics. En el següent punt veurem com s'emmagatzema i transfereix aquesta informació genètica continguda en l'ADN.

2. Genètica i epigenètica

Com em vist en el punt anterior cadascun dels processos que ens caracteritzen com a éssers vius estan definits per la informació continguda a la seqüència de nucleòtids de l'ADN. El color dels nostres ulls, la propensió a sofrir una malaltia o la intolerància a un cert tipus d'aliment es deuen a certes combinacions de nucleòtids en el conjunt gens. El conjunt de trets diferencials que mostra un organisme a nivell genètic s'anomena **genotip**, mentre que el conjunt dels seus gens, i per extensió el cúmul de la seva seqüència, s'anomena **genoma**.

En general, el genoma dels individus d'una mateixa espècie es força similar entre sí, amb variacions entorn el 0.1%, i la seva similitud decreix respecte altres espècies en funció de la distància evolutiva entre ells. Per exemple, la similitud a nivell de seqüència entre un humà i un ximpanzé es lleugerament superior al 99% però entre un humà i la mosca del vinagre trobem un 60% de similitud. Hem de pensar que la majoria dels essers vius compartim un gran nombre de processos metabòlics, des dels mecanismes de control cel·lular fins a les proteïnes estructurals que formen els diferents teixits, pel que no es estrany que la variació morfològica evident per l'ull humà no sigui tan accentuada a nivell bioquímic.

En el punt anterior, parlant del dogma central de la biologia molecular, hem comentat que l'ADN es replica a si mateix així com es transcriu en ARN. Això vol dir que qualsevol seqüència d'ADN troba el seu origen en una altra seqüència idèntica d'ADN. Aquesta **replicació** d'ADN es dona com a part del procés del **cicle cel·lular**. Al principi, tot organisme comença essent una sola cèl·lula i, almenys en organismes pluricel·lulars –eucariotes–, la replicació i diferenciació de successives generacions de cèl·lules dona lloc a organismes amb una alta complexitat funcional.

Aquesta primera cèl·lula embrionària eucariòtica es caracteritza per tenir dos jocs de gens provinents dels progenitors. A partir d'aquest punt, totes les cèl·lules no sexuals del organisme contindran aquestes dues còpies del material genètic dels progenitors. Cal comentar en aquest punt que les cèl·lules sexuals –**gàmetes**–, com per exemple els espermatozoides i òvuls en els mamífers, són un cas apart doncs només contenen una sola còpia del material genètic del progenitor que posteriorment, al unir-se, donaran lloc a les dues còpies.

Quan parlem de *material genètic* en el paràgraf anterior parlem evidentment d'ADN, però a diferencia d'abans ja no pensem en l'ADN com una estructura atòmica de doble hèlix, sinó que ara el veiem com un paquet de gens, una estructura compacta i replegada que forma els **cromosomes**.

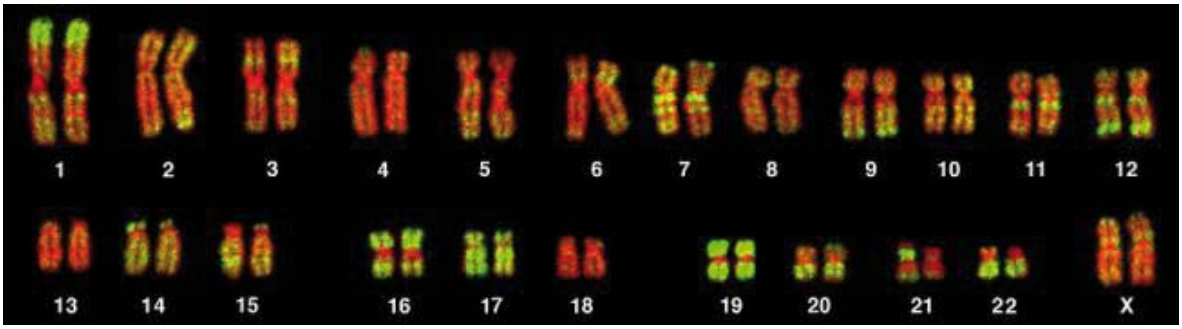


Fig. Annx. 6. Conjunt complet de cromosomes (cariotip) d'una dona. Es poden apreciar les dues còpies de cada autosoma (cromosomes no sexuals, del 1 al 22) i dues copies del cromosoma sexual (X). En el cas d'un home trobaríem els mateixos autosomes més una còpia del cromosoma X i una del cromosoma Y.

L'estructura resultant d'aquesta condensació de l'ADN formant una estructura anomenada **cromatina**. Aquesta compactació, que permet emmagatzemar fins a 2 metres d'ADN dins el nucli d'una cèl·lula d'uns pocs nanometres –una mil millonèsima part d'un metre– de diàmetre es possible, en part, gràcies a unes proteïnes anomenades **histones**, les quals, al unir-se a l'ADN formen el que es coneix com a **nucleosomes.** L'organització i els efectes d'aquesta organització en els nucleosomes és el cos central d'estudi d'aquesta tesi i la seva descripció formal es troba en la introducció d'aquesta memòria.

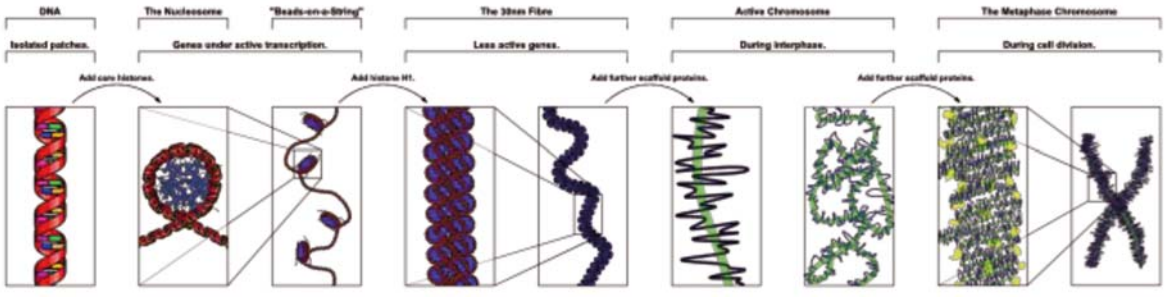


Fig. Annx. 7. Procés de compactació de l'ADN. El DNA (quadre de l'esquerra) es compacta en un primer nivell al voltant dels nucleosomes (segon quadre), fins arribar a l'estructura del cromosoma (dreta).

Amb el que hem vist fins aquí, semblaria que l'única variabilitat que podem trobar a nivell genètic vindria donada per la presència de diferents pares, però això no explicaria perquè els organismes poden evolucionar de forma diferent afavorint la seva especialització i supervivència. A pesar de que la vida requereix que la replicació de l'ADN sigui perfecta, donada la complexitat del procés això no sempre és així. Hem de pensar que al llarg de la nostra vida repliquem milions i milions de vegades les nostres seqüències d'ADN i que és estadísticament probable que en algunes d'aquestes copies es produeixi un error.

Aquests *errors*, conjuntament amb altres factors, provoquen mutacions aleatòries a la nostra cadena d'ADN. Algunes mutacions no tenen major transcendència doncs passen en regions no codificants del genoma. Altres mutacions produeixen alteracions en la seqüència d'aminoàcids d'una proteïna i fan la cèl·lula inviable, pel que és mor i aquesta mutació no es transmet. En altres casos la pròpia maquinària cel·lular detecta un error en la replicació del ADN i dona una senyal que desencadena la mort cel·lular –apoptosis– en pro del organisme. Però en un petit nombre de vegades, aquestes mutacions escapen dels controls, i permeten la viabilitat de la cèl·lula. Llavors ens podem trobar amb cèl·lules que afectin a la supervivència del individu o que, per el contrari, afavoreixin la seva supervivència. Segons la teoria de la selecció natural proposada per Darwin, les primeres mutacions no es propagaran doncs mataran a l'individu, però les segones li permetran una major adaptació al medi i per tant seran transmeses amb més facilitat les generacions següents.

Quan ens trobem amb un gen que presenta mutacions a nivell de la seqüència parlem de diferents **al·lels**. En aquest cas ens trobem amb un gen que potencialment pot fer la mateixa funció, però amb lleugers canvis. Per exemple, un canvi en la seqüència d'una proteïna amb una funció enzimàtica pot afectar en l'eficiència amb que fa la seva funció, permetent-nos una assimilació més ràpida o més lenta d'un determinat fàrmac o aliment.

Hem anomenat *genotip* al conjunt de gens present en un individu donat, però ara hem vist com un mateix gen pot originar trets diferents en funció de diferents factors. La majoria de la població humana té el gen de la tirosinasa, el qual és fonamental en la producció de la melanina, un pigment natural present en vàries de les nostres cèl·lules i que, conjuntament amb altres proteïnes, afecta el color de la pell, els cabells o els ulls. La manca o malformació d'aquest gen origina l'albinisme (individus sense pigments), però que tinguem el gen no vol dir que tots tinguem el mateix color de pell o cabell. Quan veiem una persona típicament nòrdica, amb els cabells rossos i els ulls blaus, i una per-

sona sud-americana amb els cabells morens i els ulls foscos estem observant dos trets diferents relacionats amb la pigmentació de la pell a pesar de que tots dos comparteixen el mateix genotip. Aquesta expressió diferencial del tret l'anomenem **fenotip**, que es pot definir com l'expressió del genotip modulada per la interacció amb el medi.

El concepte de fenotip permet introduir un concepte com el de **regulació genètica**. El simple fet de tenir en el nostre ADN una seqüència donada codificant per proteïnes no vol dir que sempre estiguem generant –expressant– aquesta proteïna, ni que sempre ho fem amb la mateixa quantitat. De fet, les cèl·lules de diferents teixits del nostre organisme tenen patrons diferencials d'expressió que els hi permeten exercir funcions diferents a pesar de tenir el mateix genoma.

Els elements de regulació poden ser de naturalesa molt variada i seria terriblement complex abordar-los tots en aquesta introducció. Tot i així, si que els podríem agrupar en 3 tipus en funció del nivell sobre el que actuen les seves modificacions: estructural, post-transcripcional i post-transduccional.

Les modificacions **estructurals** fan referència a modificacions sobre l'estructura de l'ADN o la cromatina canviant una base per una altra, introduint alguna variant en un nucleòtid (com podria ser la metilació de la citosina) o modificant parts de les histones en els nucleosomes. Els efectes d'aquestes modificacions es poden materialitzar en canvis a l'accessibilitat de l'ADN, canvis a la seva seqüència o canvis en diferents marques que poden permetre o impedir el reconeixement per part de la maquinària de traducció.

Les modificacions **post-transcripcionals** afecten l'ARN missatger, per exemple interferint en la seva interacció amb el ribosoma i modulant així la quantitat generada d'una determinada proteïna.

Finalment, les modificacions **post-transduccionals** afecten a l'estructura o plegament d'una proteïna, afectant la seva funcionalitat o la seva interacció amb altres proteïnes.

En el cas d'aquesta tesi es parla sobretot de modificacions estructurals, sobretot les referents al paper regulador de les histones i la metilació de l'ADN (Fig. Annx. 8).

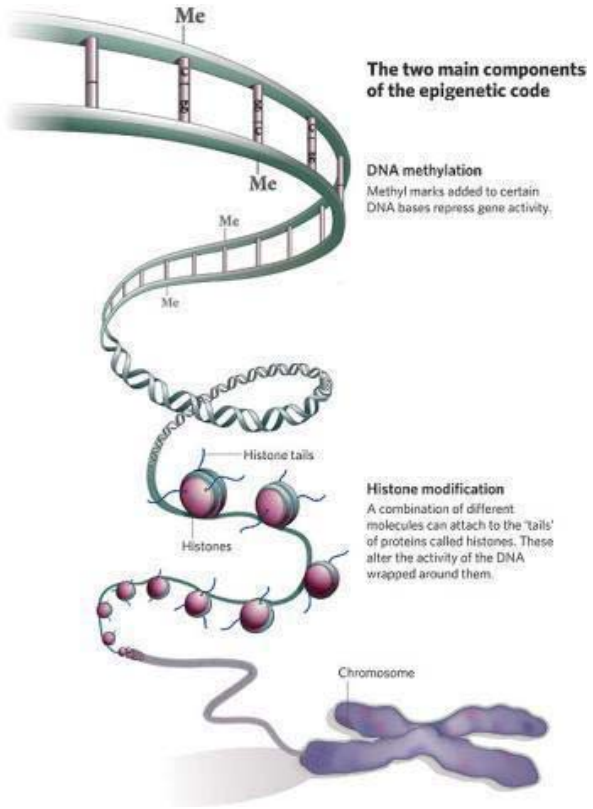


Fig. Annx. 8. Esquema de l'organització dels cromosomes en nucleosomes i les seves dues principals modificacions epigenètiques, la metilació de l'ADN i la modificació d'histones. (font: Nature 441)

Pensem, per exemple, amb dos bessons monocigòtics que comparteixen la totalitat del seu genoma. Si els dos germans viuen estils de vida diferents (per exemple, un fuma i l'altre no) aquestes diferencies poden provocar canvis en la regulació de determinats gens, que eventualment es poden traduir en diferents mecanismes de control més enllà de la pròpia seqüència d'ADN.

Com es pot veure, la regulació genètica es un complex procés on diferents modificacions i la seva interacció bidireccional entre la pròpia dinàmica de les proteïnes i el medi acaba definit el funcionament d'una cèl·lula de l'organisme en un moment donat. A mesura que el coneixement sobre el genoma humà ha anat creixent, hem observat que no podem atribuir només a la seqüència genètica tota la informació que rebem dels progenitors. Les modificacions estructurals sobre l'ADN i la cromatina en general tenen la propietat de mantenir-se parcialment i actuar com a reguladors estructurals en noves cèl·lules, incloent això a les cèl·lules embrionàries, és a dir, als descendents. Per tant, no podem pensar l'herència genètica només a nivell de seqüència del genoma. Tota aquesta informació rellevant per les cèl·lules per sobre de la pròpia genètica rep el nom d'**epigenètica**.

ANNEX 2: SUPPLEMENTARY MATERIALS

1. Physical properties of naked DNA influence nucleosome positioning and correlate with transcription start and termination sites in yeast

Supplementary material for this work includes:

- Additional Methods
- Additional Figures
- Additional Tables

Additional Methods

Naked DNA preparation

Cultures of *Saccharomyces cerevisiae* strain BY4741 were grown for 20 h shaking at 30 ^{o}C in YPD rich medium. The cells were harvested by centrifugation and resuspended in a buffer containing 50 mM EDTA and 0.5 g L^{-1} zymolase (Seigaku, Inc.) to generate spheroplasts, which were pelleted and incubated at 65 ºC for 30 min with Tris-EDTA (TE) buffer (50 mM Tris pH 7.4, 20 mM EDTA) and 1% (wt/vol) SDS. Subsequently, SDS-protein complexes and chromosomal DNA were precipitated by adding potassium acetate and collected by centrifugation. The DNA-containing pellets were washed with absolute ethanol and resuspended in TE buffer (10 mM Tris pH 8, 1mM EDTA). Samples were treated with 0.08 g L^{-1} DNase-free RNase for 1 h at 37 ^{o}C and purified by phenol:chloroform extraction and ethanol precipitation. Purified DNA samples were quantified by *Qubit* fluorometer (Invitrogen, Inc.) and *Nanodrop* spectrophotometer (Thermo Scientific, Inc.).

Digestion of naked DNA

Naked DNA samples were fragmented either by MNase digestion or *Bioruptor* disruption. For MNase digestion, samples containing 20 µg of naked DNA were digested at 28 ^{o}C for 5 min with micrococcal nuclease (Sigma-Aldrich, Inc.) at concentrations of 0, 0.01, 0.03, 0.06 and 0.1 U, respectively. The digestion reactions were then quenched with 10 mM EDTA and purified by ethanol precipitation. Fragmentation by *Bioruptor* system was performed with 5 µg of DNA sonicated during 0, 5, 10 and 15 minutes (at intervals of 10 s on-30 s off), respectively. In both approaches, the purified samples were examined by 2% agarose gels (**Additional Figure A8A**) and the reactions containing a fragment size of 100-350 bp were selected for DNA sequencing.

Nucleosomal DNA preparation

Overnight cultures of *Saccharomyces cerevisiae* strain BY4741 were diluted to an OD_{600} of 0.2 using fresh YPD media and further grown at 30 ^{o}C until reaching an OD_{600} of 0.8–0.9. Cells were cross-linked with 2% (v/v) formaldehyde for 30 min while shaking at 30 ^{o}C and then the reaction was stopped by the addition of 125 mM glycine. Cells were harvested, washed with PBS buffer and resuspended in 1 M sorbitol in 50 mM Tris pH 7.4 with freshly added 10 mM β-mercaptoethanol. Subsequently, zymolase

was added to a final concentration of 0.25 g L^{-1}. Cells were spheroplasted at 30 ^{o}C for 40 min, pelleted and resuspended in 1 M sorbitol, 50 mM NaCl, 10 mM Tris pH 7.4, 5 mM $MgCl_2$, 1 mM $CaCl_2$ and 0.075% (vol/vol) Nonidet P40, with freshly added 1 mM β-mercaptoethanol and 500 μM spermidine. Different digestion reactions were setup with MNase at concentrations of 0, 0.04, 0.1, 0.15 and 0.3 U, respectively. The digestion reactions were incubated at 37 ^{o}C for 30 min and stopped by adding 20 mM EDTA. To remove the protein contents, 0.8 g L^{-1} proteinase K was added and incubated overnight at 65 °C. DNA samples were treated with DNase-free RNase (1 g L^{-1}) for 1 h at 37 ^{o}C and purified by phenol:chloroform extraction and ethanol precipitation. DNA fragments were examined by 2% agarose gels (**Additional Figure A8B**). Those reactions containing at least 90% mononucleosomal DNA fragments were selected for sequencing.

Generation of high-throughput sequencing reads

High-throughput sequencing reads in *qfasta* format were obtained from a sequencing facility using Illumina Genome Analyzer (GA) IIx. MNase-digested DNA samples were sequenced in 54 cycles with 7 extra cycles for multiplex indexing from both ends and subsequently pre-processed with standard Illumina GA base-call pipeline using ELAND 1.5.1 and CASAVA 1.7 software. MNase digested experiments were carried out in duplicates. A strong mean correlation was observed among genome-wide reads coverage (Spearman correlation ρ=0.770), that increased till 0.982 or 0.997 in the window of 1000 bp surrounding the TSSs or TTSs, respectively. Raw data is available through the NBCI Short Read Archive under accession number SRA030453.

Reads alignment and pre-processing

Reads were aligned on a self-compiled index of the October 2003 *Saccharomyces cerevisae* genome using the Bowtie algorithm [1]. Genome sequences were obtained from the UCSC genome browser (http://hgdownload.cse.ucsc.edu/goldenPath/sacCer1 /bigZips/, date of access: 2010-February-05). Reads were mapped allowing a maximum of three mismatches and an insert length of 500 bp. Due to the presence of repetitive sequences along the genome, those reads that could be aligned to multiple regions were mapped to all the possible places, avoiding depleted region artifacts. This procedure allowed a reliable detection of abnormally low-coverage regions, but it is not informative of the coverage in the repeated regions of the genome. In this way, false depleted regions due the mapping process were removed. On average, 92.77% of the

reads reported at least one alignment. For the nucleosome calling, the same reads have been mapped again by accounting only for uniquely-mapping reads to avoid artificially read-enriched regions. The coverage obtained from these uniquely-mapping reads has been processed to locate nucleosome as described below (see also main Methods section).

Reads importation and duplicate reads removal

In all cases, reads described in the previous point were imported to R/Bioconductor framework[2]. Reads were analyzed and processed with htSeqTools[3] package for quality control and over-amplification correction. Reads with a probability higher than 95% of being over-amplification artifacts were removed.

Reads coverage calculation

For each sample, the MNase degradation profiles were calculated as the number of reads at every position across the genome. The coverage was normalized independently from the total number of reads generated in every sequencing dataset. For each experiment, we removed duplicated reads and divided the coverage value by the total number of reads. Normalized coverage values were subsequently scaled by a factor of 10^6 to reads per million (r.p.m.). To improve the visual identification of the nucleosome dyads, and only for visualization purposes (the rest of the study was performed accounting the entire reads), the coverage of a nucleosomal sample was calculated by trimming each single read of nucleosomal DNA to its middle 40 bp, around the dyad (example in **Additional Figure A9**).

Identification of Low coverage Regions (LRs)

The identification of LRs, i.e. genomic segments with non-zero coverage below a certain coverage percentile, was performed on naked and nucleosomal DNA reads coverage maps. For naked DNA, we defined a LR as a region within the lower 2.5 percentile of the sample. This threshold was determined by manual inspection of the coverage maps in the genome (example in **Additional Figure A10**). Accordingly to the nature of the experiment and our mapping procedure (see above for details), zero-coverage regions were attributed to experimental and processing artifacts and were not selected. For nucleosomal DNA a less restrictive threshold (10 percentile) was used to take into account the larger amount of depleted regions intrinsically caused by the

nucleosome free regions. This procedure guarantees similar read counts for naked and nucleosomal DNA samples (see **Additional Figure A11)**.

Neighboring LRs in a distance shorter than 4 bp were merged, and LRs shorter than 5 bp were removed. Very long regions (typically regions larger than ~250 bp with a size up to the 95 percentile of the LR length distribution) were discarded to avoid artifacts caused by errors in sequencing or mapping. Common low coverage regions (CLRs) were identified as the intersection of LRs present in both naked and nucleosomal samples. LRs lengths are between 5 and 250bps with an average length of 50bps

To discard possible artifacts in the identification of LRs we compared the coverage maps of naked DNA fragmented both by MNase and sonication, without detecting any substantial bias on the LRs marked by MNase digestion. The genome-wide identification of LRs in naked DNA fragmented using sonication resulted in 84 regions (1,394 bp), indicating that LRs observed in naked DNA treated with MNase were not affected by sequencing artifacts. (**Additional Figure A1**).

Identification and characterization of Common Low Regions (CLRs)

We defined common low regions (CLRs) as the base pair-wise intersection of LRs from paired-end sequenced samples for both nucleosomal and naked DNA. 2,770 regions were identified (139,285 bp: 57.60% of the LRs in MNase-digested naked DNA). Every CLR across the genome was located respect to the nearest TSSs by calculating the minimum absolute distance from both 3' and 5' ends of the CLR to the closest TSS. The distance was considered as negative if CLRs were upstream of TSSs, and positive otherwise. In case of two equidistant TSSs, upstream (negative) values were chosen. The same procedure was applied to locate CLRs respect to the nearest TTSs. The manipulation and intersection of the regions were performed with R/Bioconductor[2]

MNase-preferred cut sites and tetramer composition of degraded regions

MNase cut sites were extracted (after statistical duplicate reads removal as described above) by taking the tetramer composed of the two bases upstream and two bases downstream of each read end.

The tetramer composition analysis of (C)LRs considered all the overlapping tetramers in the selected regions. The frequencies of complementary tetramers in reverse strands were summed up to account for symmetrical structure of DNA. To calculate the expected tetramer frequency, ten million tetramers were sampled in the entire yeast

genome. The ratio between the experimentally observed and expected tetramer frequency was calculated and used to point out a possible over- or under-representation. The significance (p-value) of the enrichment or depletion was calculated for ten million random observations. P-value for over-represented tetramers (ratio > 1) was calculated as the fraction of times (10,000 observations) that the frequency, observed in a population of 1,000 randomly selected tetramers from the genome, was equal or smaller than the expected frequency. P-value for under-represented tetramers (ratio < 1) was calculated in a similar way but counting the number of times that the tetramer was found in a greater frequency than the expected.

Nucleosome calling and MNase bias correction

A new peak detection algorithm has been released recently to locate nucleosome dyads from read coverage values[4]. The peaks in the coverage maps identified the dyads, while the surrounding 74 bases on both sides determined the location of the nucleosomes. Peak detection algorithm was refined to only consider enriched regions in the coverage map. These regions were further smoothed using Fourier-Analysis, applying a standard noise filter based on principal component selection and signal reconstitution. Spearman correlation between original and noise-filtered coverage maps is 0.97. Resulting nucleosome calls were scored considering the width and the height of the coverage peaks and then classified as non-overlapping calls ("well-positioned" nucleosomes) or as overlapping calls ("fuzzy" nucleosomes). All the steps described here are implemented and explained in *nucleR* package[4] documentation. **Additional Figure A9** shows in detail nucleosome call maps.

Correction of nucleosomal digestion profiles was done by assuming that a randomly distributed coverage map of naked DNA reads will cause a uniform coverage profile (confirmed by the sonication experiments; see main **Figure 3**). We set this profile to a constant value equal to the mean value of naked DNA coverage. We compare this uniform profile with the experimentally obtained for the naked DNA. We assumed that deviations from the background quantify sequence-dependent MNase biases, which were then used to correct nucleosomal DNA coverage profile. The method for this processing is documented in nucleR library in "*controlCorrection*" function[4].

Physical descriptors and nucleosome deformation energy

As described in detail elsewhere [5,6], we collected equilibrium MD trajectories (150 ns long; T=298 K, P=1 atm.) in water (more than 9,000 TIP3P molecules Na^+ as

counterion) using state of the art simulation protocols for four duplexes, which contain the ten unique dinucleotide steps (steps (d(GG)·d(CC), d(GC)·d(GC), d(GA)·d(YC), d(GT)·d(A·C), d(AG)·d(CT), d(AA)·d(TT), d(AT)·d(AT), d(CG)·d(CG), d(CA)·d(TG) and d(TA)·d(TA)): d((GCCTATAAACGCCTATAA)·d(TTATAGGCGTTTATAGGC), d(CTAGGTGGATGACTCATT)·d(AATGAGTCATCCACCTAG), d(CACGGAACCGGTTCCGTC)·d(GACGGAACCGGTTCCGTG) and d(GGCGCGCACCACGCGCGG)·d(CCGCGCGTGGTGCGCGCC). Trajectories were projected into helical space to determine the covariance matrix at each step, from which stiffness parameters were obtained:

$$\Theta = k_B T C^{-1} = \begin{pmatrix} k_w & k_{wr} & k_{wt} & k_{ws} & k_{wl} & k_{wf} \\ k_{wr} & k_r & k_{rt} & k_{rs} & k_{rl} & k_{rf} \\ k_{wt} & k_{rt} & k_t & k_{st} & k_{tl} & k_{tf} \\ k_{ws} & k_{rs} & k_{st} & k_s & k_{ls} & k_{lf} \\ k_{wl} & k_{rl} & k_{tl} & k_{ls} & k_l & k_{lf} \\ k_{wf} & k_{rf} & k_{tf} & k_{lf} & k_{lf} & k_f \end{pmatrix}$$

where k_b is the Boltzman constant, T is the absolute temperature, and k stands for the different stiffness constants defining the 36 elements of the stiffness matrix (Θ) (twist (w), roll (r), tilt (t), rise (s), slide (l) and shift (f)) at the dinucleotide level obtained by inversion of the MD-associated covariance matrix (C). In order to test whether nearest-neighboring effects were important for our purposes, we repeated the stiffness analysis using trajectories for all 136 unique tetramers extracted from the Ascona B-DNA Consortium (ABC) [7]. As noted in the main text, the results obtained when using tetramer resolution level were identical to those derived when nearest-neighboring effects were neglected. This finding supports the robustness of our calculations and strongly suggests that, for average-whole genome analysis as the one reported here, nearest-neighboring corrections have small impact.

Stiffness matrix described above was also used to determine *ab initio* (i.e. without any knowledge-based training) the energy required to wrap a 147 bp long DNA sequence into a nucleosome conformation, assuming that distortion is naturally harmonic. This was determined as:

$$E = \frac{1}{2}\Theta(X - X_0)^2$$

where X stands for the (helical) geometry of the DNA in the crystal structure of nucleosome, and X_0 stands for the equilibrium geometry of the same sequence of DNA in water in the absence of histones (also obtained from MD). The reference nucleosome structure was obtained by averaging and smoothing of all available X-ray structures of the nucleosome core particle [8-15] using a Fourier Transform algorithm [16]. This procedure reduces local variability that can be due to crystallization artifacts. Note that large E values signal those regions where physical descriptors indicate that wrapping a DNA in a left-handed superhelix is expected to be difficult.

References

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

2. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JYH, Zhang J: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome biology* 2004, **5**:R8010.1186/gb-2004-5-10-r80.

3. Planet E, Stephan-Otto C, Reina O, Flores O, Rossell D: **htSeqTools: High-Throughput Sequencing Quality Control, Processing and Visualization in R**. *Submitted (available in Bioconductor repository)* 2011, **-**:-.

4. Flores O, Orozco M: **nucleR: a package for non-parametric nucleosome positioning.** *Bioinformatics (Oxford, England)* 2011, **27**:2149-215010.1093/bioinformatics/btr345.

5. Faustino I, Pérez A, Orozco M: **Toward a consensus view of duplex RNA flexibility.** *Biophysical journal* 2010, **99**:1876-85.

6. Pérez A, Lankas F, Luque FJ, Orozco M: **Towards a molecular dynamics consensus view of B-DNA flexibility.** *Nucleic acids research* 2008, **36**:2379-94.

7. Lavery R, Zakrzewska K, Beveridge D, Bishop TC, Case DA, Cheatham T, Dixit S, Jayaram B, Lankas F, Laughton C, Maddocks JH, 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 research* 2010, **38**:299-31310.1093/nar/gkp834.

8. Luger K, Mäder AW, Richmond RK, Sargent DF, Richmond TJ: **Crystal structure of the nucleosome core particle at 2.8 A resolution.** *Nature* 1997, **389**:251-6010.1038/38444.

9. Harp JM, Hanson BL, Timm DE, Bunick GJ: **Asymmetries in the nucleosome core particle at 2.5 A resolution.** *Acta crystallographica. Section D, Biological crystallography* 2000, **56**:1513-34.

10. Suto RK, Clarkson MJ, Tremethick DJ, Luger K: **Crystal structure of a nucleosome core particle containing the variant histone H2A.Z.** *Nature structural biology* 2000, **7**:1121-4.

11. Suto RK, Edayathumangalam RS, White CL, Melander C, Gottesfeld JM, Dervan PB, Luger K: **Crystal structures of nucleosome core particles in complex with**

minor groove DNA-binding ligands. *Journal of molecular biology* 2003, **326**:371-80.

12. Davey CA, Sargent DF, Luger K, Maeder AW, Richmond TJ: **Solvent Mediated Interactions in the Structure of the Nucleosome Core Particle at 1.9Å Resolution†**. *Journal of Molecular Biology* 2002, **319**:1097-1113.

13. Muthurajan UM, Bao Y, Forsberg LJ, Edayathumangalam RS, Dyer PN, White CL, Luger K: **Crystal structures of histone Sin mutant nucleosomes reveal altered protein-DNA interactions.** *The EMBO journal* 2004, **23**:260-71.

14. Ong MS, Richmond TJ, Davey CA: **DNA stretching and extreme kinking in the nucleosome core.** *Journal of molecular biology* 2007, **368**:1067-74.

15. Bao Y, White CL, Luger K: **Nucleosome core particles containing a poly(dA.dT) sequence element exhibit a locally distorted DNA structure.** *Journal of molecular biology* 2006, **361**:617-24.

16. Lavery R, Moakher M, Maddocks JH, Petkeviciute D, Zakrzewska K: **Conformational analysis of nucleic acids revisited: Curves+.** *Nucleic acids research* 2009, **37**:5917-29.

17. Mavrich TN, Jiang C, Ioshikhes IP, Li X, Venters BJ, Zanton SJ, Tomsho LP, Qi J, Glaser RL, Schuster SC, Gilmour DS, Albert I, Pugh BF: **Nucleosome organization in the Drosophila genome.** *Nature* 2008, **453**:358-6210.1038/nature06929.

Additional figures

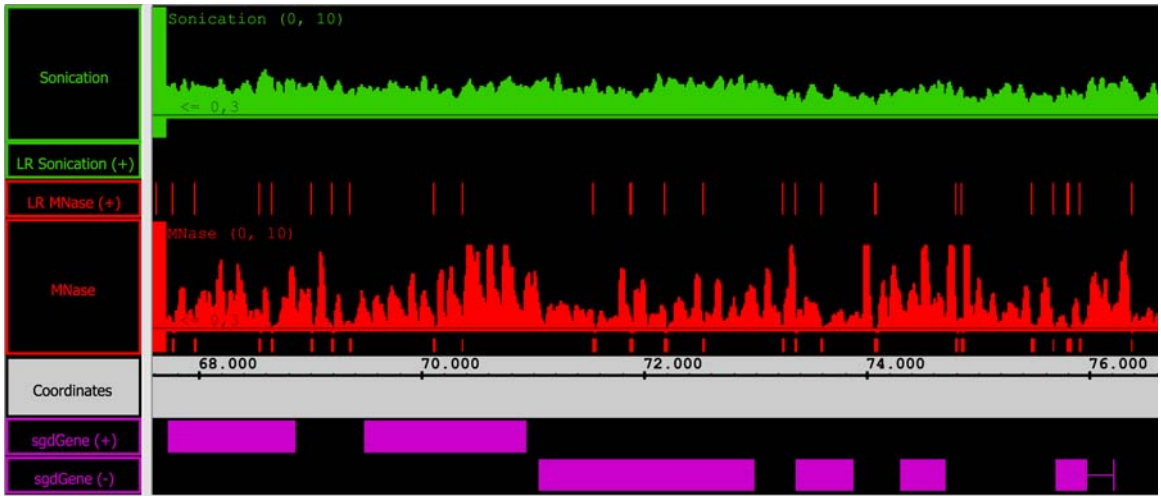


Figure A1. Coverage profiles

Coverage profile and identification of low regions (shown as vertical lines; threshold 2.5%) in chromosome 16 for sonicated naked DNA (up, green) and for MNase-digested naked DNA (bottom, red). Sonication coverage does not show any evident low region. Coordinates of chromosome 16 and both strand genes are displayed at the bottom.

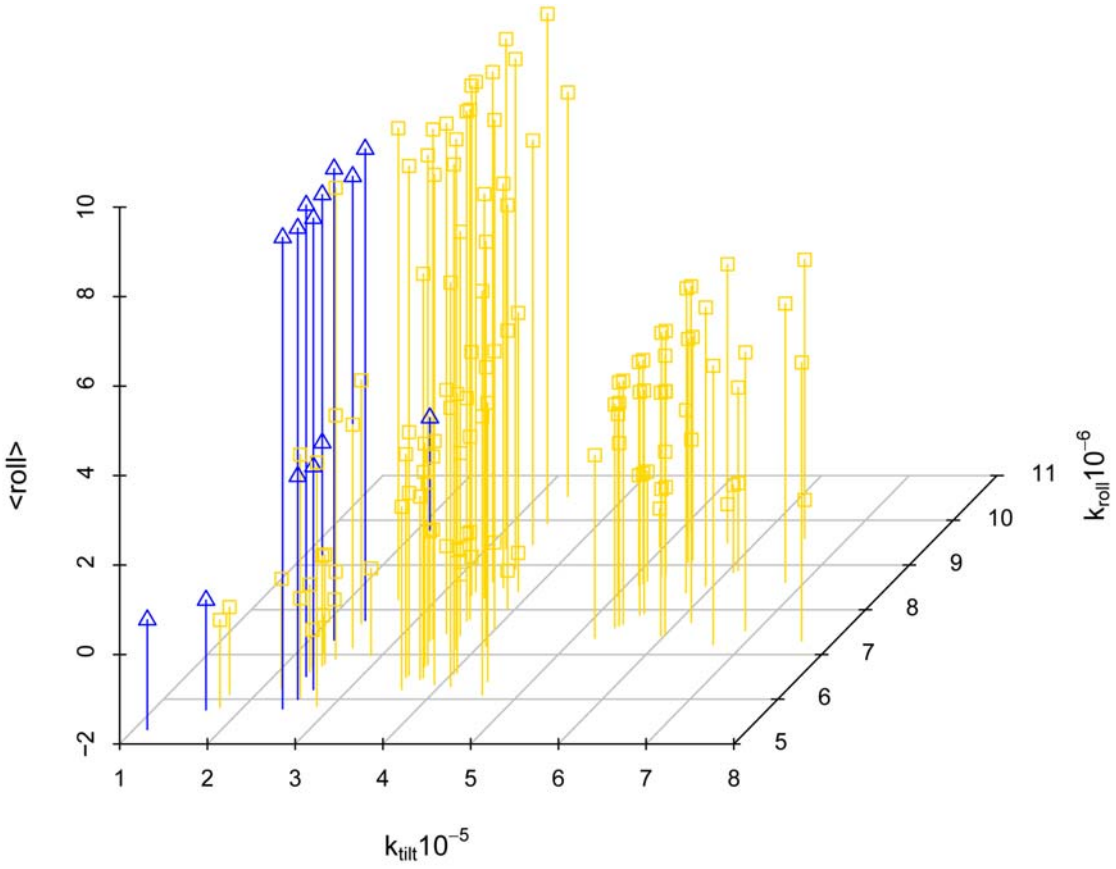


Figure A2. MNase-preferred cutting sites and physical properties

Representation of preferential (blue-triangles) vs. non-preferential (yellow-squares) MNase cutting sites in naked DNA with respect to physical properties tilt and roll stiffness (in $kcalmol^{-1}degree^{-2}$), and equilibrium roll (in degrees) for each tetramer.

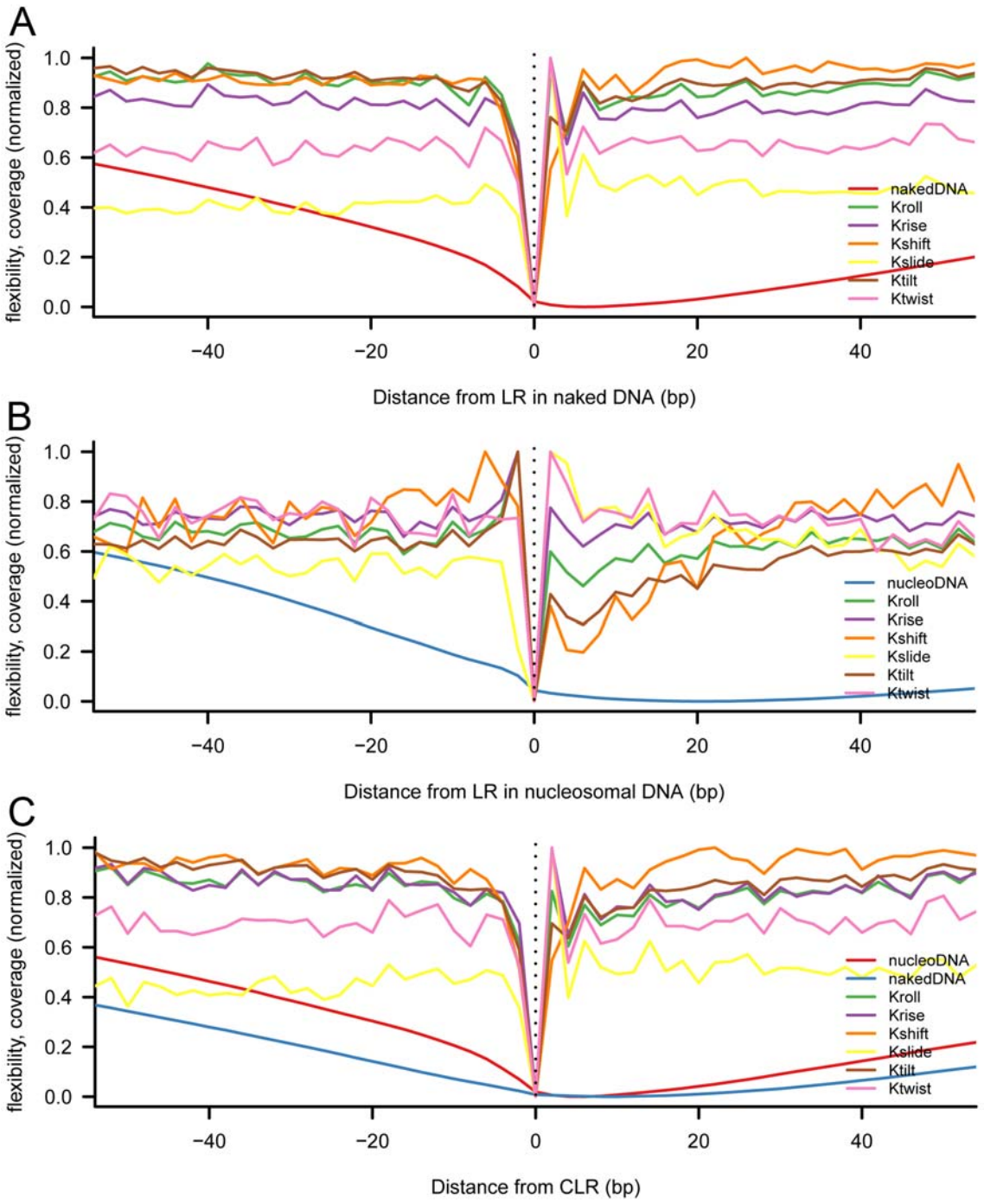


Figure A3. Individual Stiffness profiles in low coverage regions

Six individual stiffness parameters (k_{roll}, k_{tilt}, k_{twist}, k_{shift}, k_{rise} and k_{slide}) and coverage maps were calculated and averaged across all yeast genome, around (A) LRs in naked DNA, (B) LRs in nucleosomal DNA and (C) CLRs in nucleosomal and naked DNA. All values are normalized (in the 0-1 range) to facilitate analysis and comparisons.

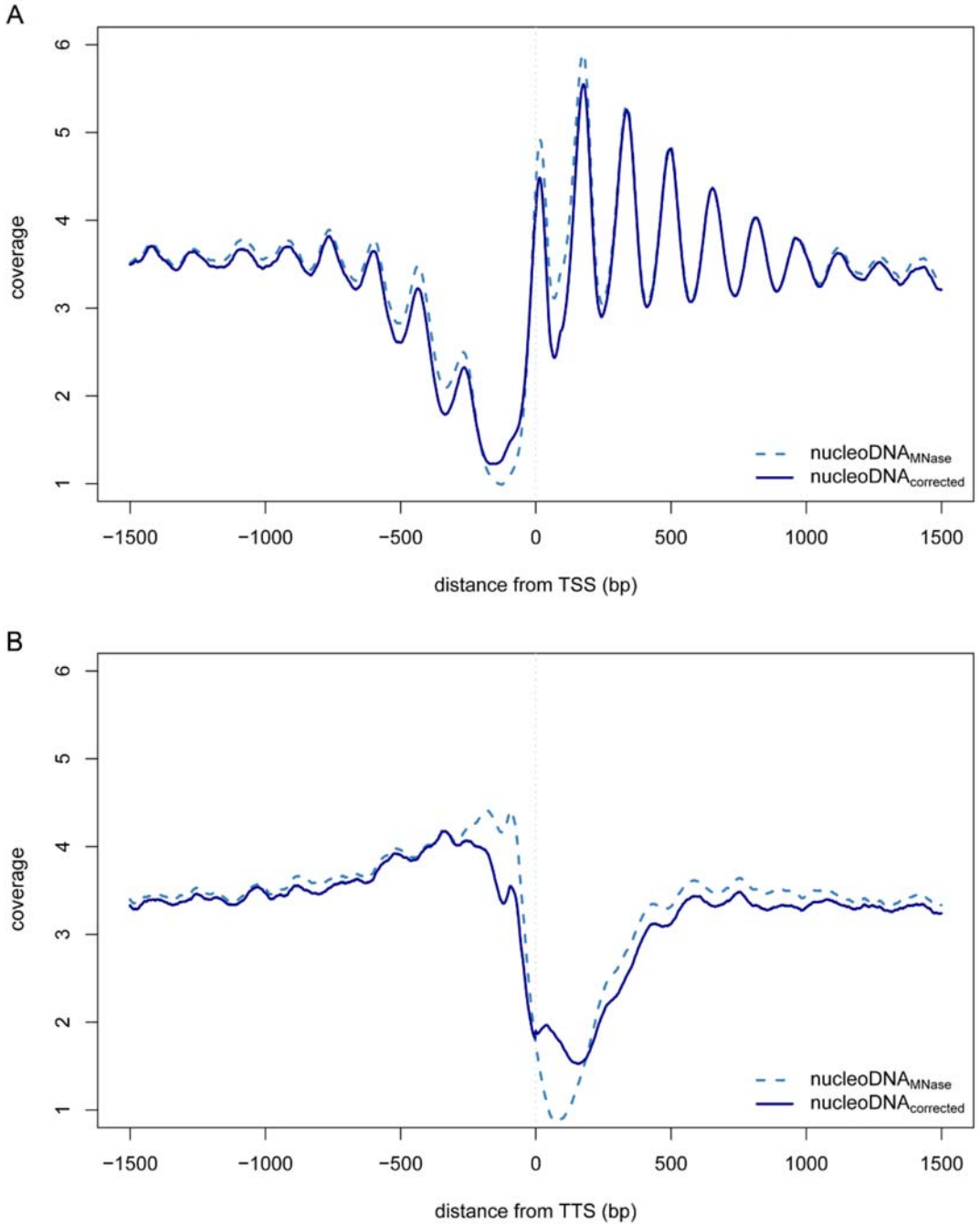


Figure A4. Average TSS and TTS coverage profiles

Coverage profiles at transcription start sites (TSSs) (top) and transcription termination sites (TTSs) (bottom) in MNase-digested nucleosomal DNA before (dashed lines) and after naked DNA correction (continuous lines). Average of 5,750 selected genes.

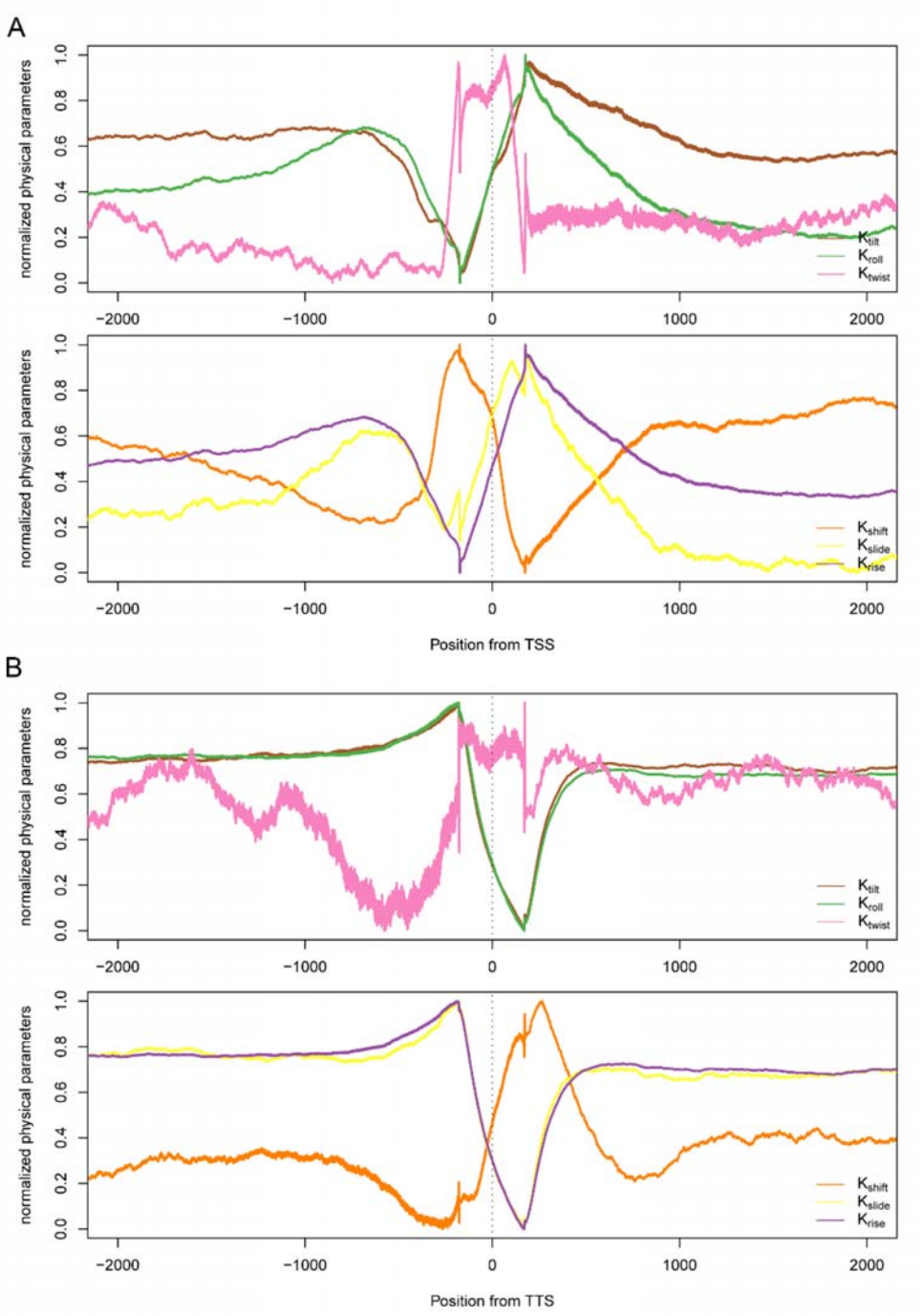


Figure A5. Variation of stiffness descriptors

Plots showing the average variation of stiffness parameters (translational or rotational) around TSSs and TTSs in the yeast genome (5,750 genes were considered).

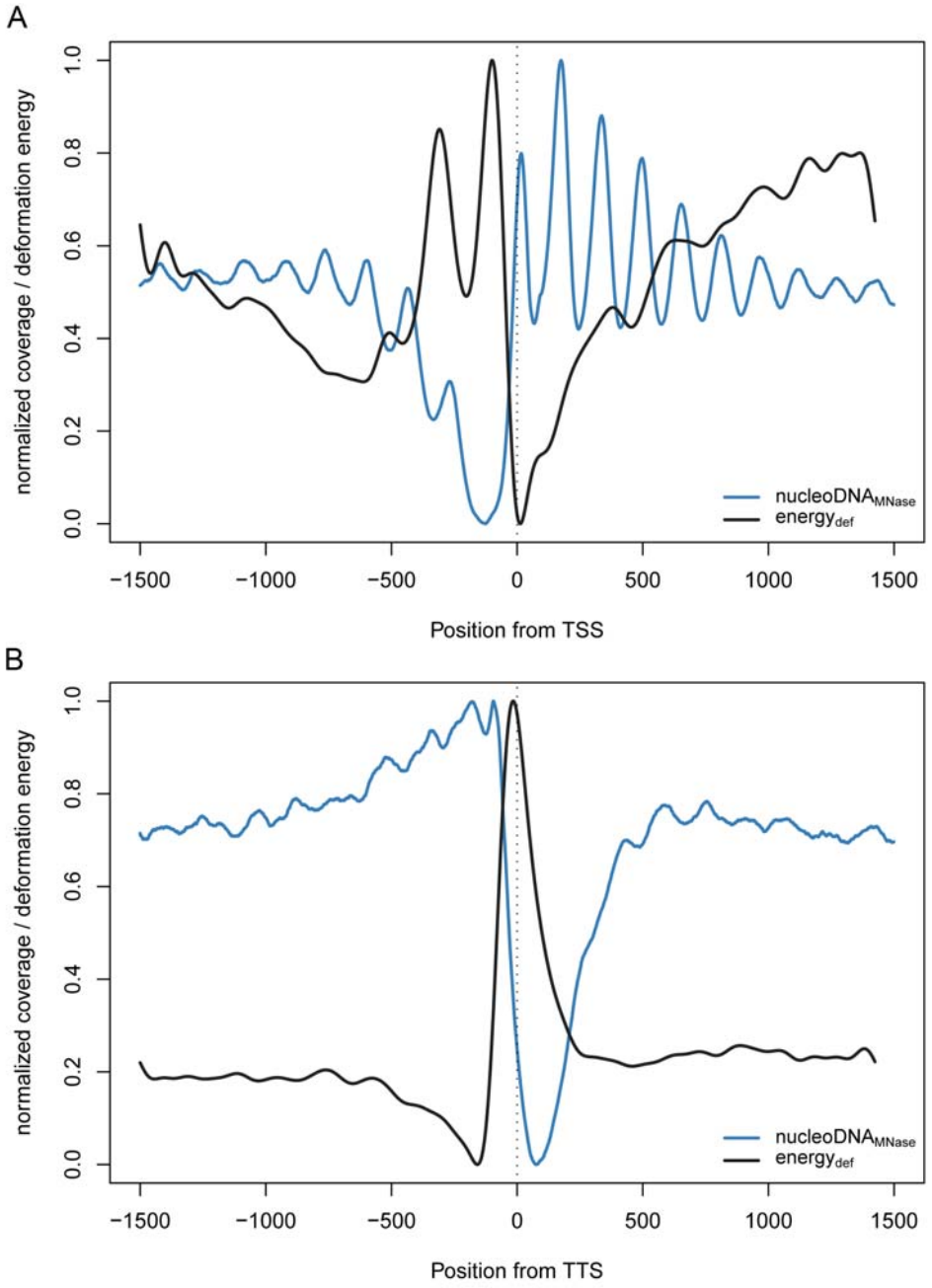


Figure A6. Nucleosome deformation energy

Plots displaying the average nucleosome deformation energy for sites around TSSs (top) and TTSs (bottom). Larger values indicate higher difficulty of DNA wrapping around a histone core. In both cases, the coverage MNase-digested nucleosomal DNA is shown as a reference. This average is calculated over 5,750 genes.

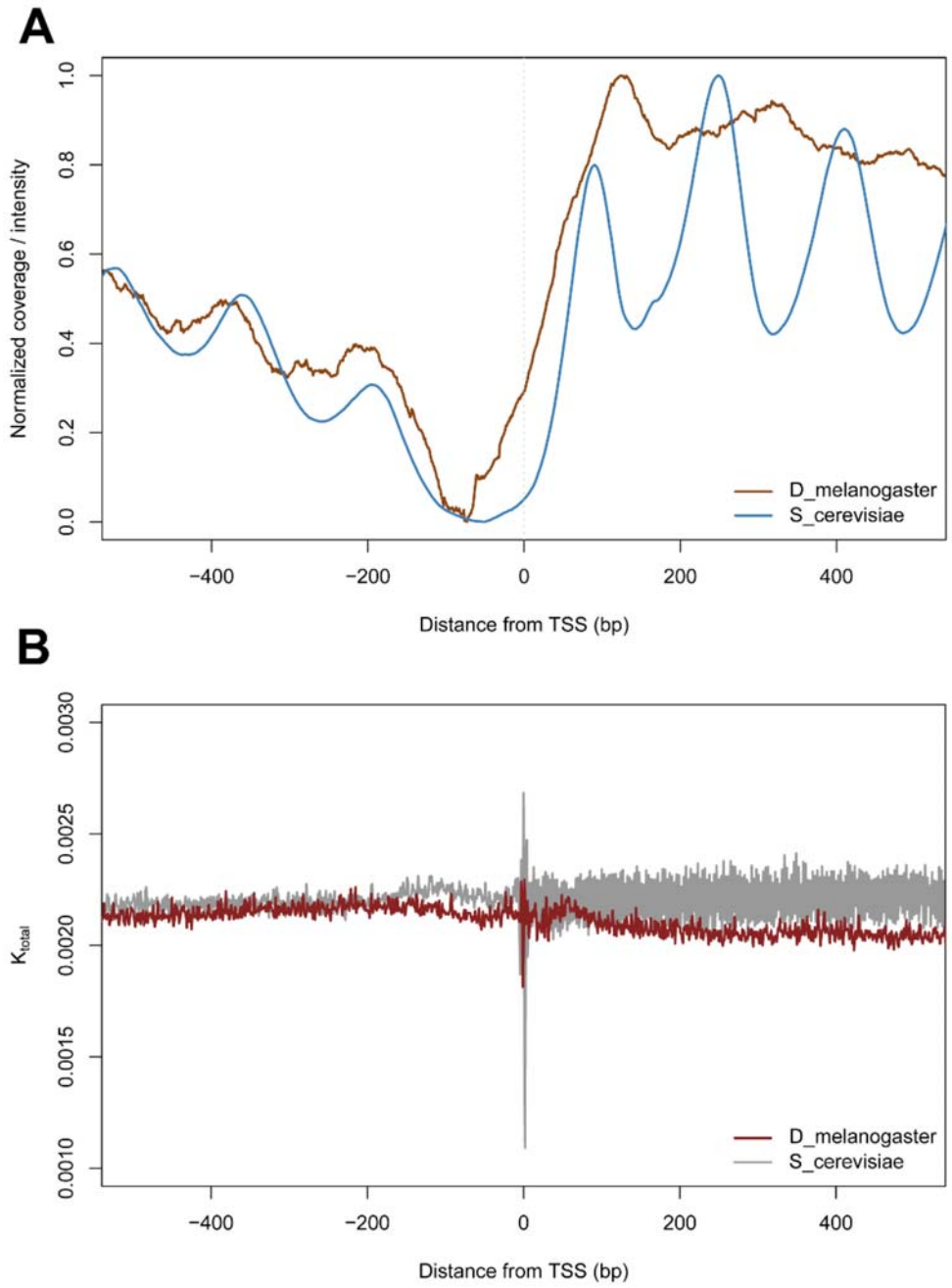


Figure A7. Comparison of coverage and stiffness profiles at TSSs in Drosophila and Saccharomyces genomes

(a) Coverage maps for nucleosomal DNA in yeast and fly[17] genomes at TSSs. Values have been normalized to account for different sequencing depths.

(b) Total stiffness parameter (k_{total}) calculated and averaged across all yeast and fly genomes around TSSs.

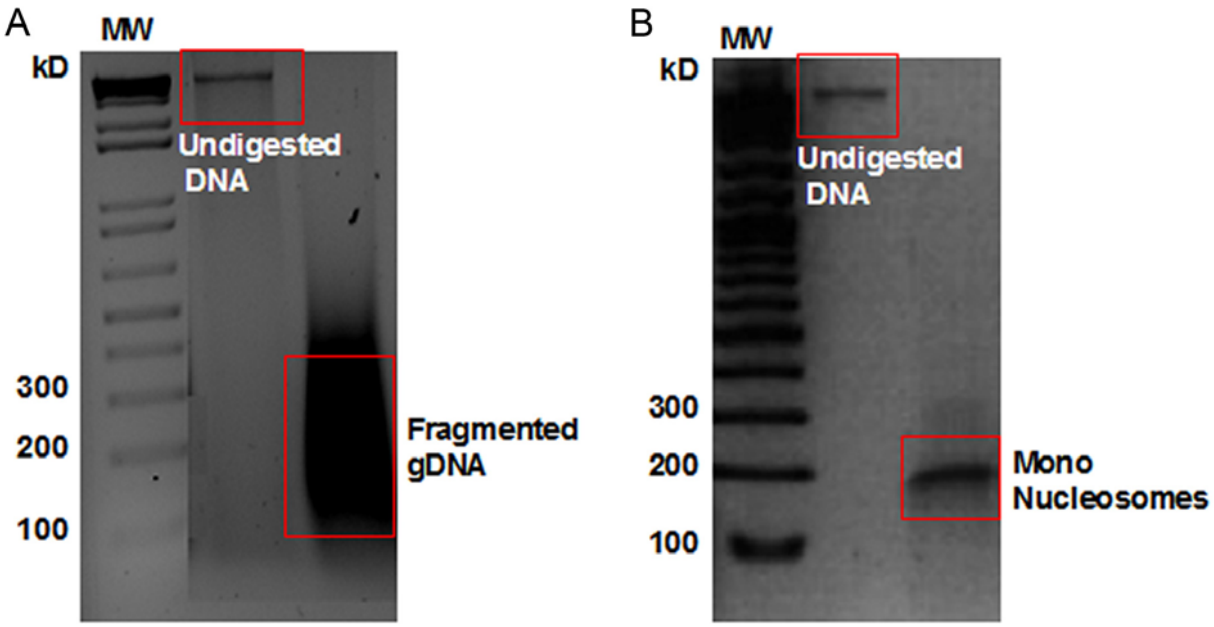


Figure A8. DNA sample purity

Native 2% agarose gels showing the genomic (A) and chromatin DNA (B) digestion products before and after MNase treatment, respectively. Fragment sizes were estimated according to standard DNA molecular markers (MW).

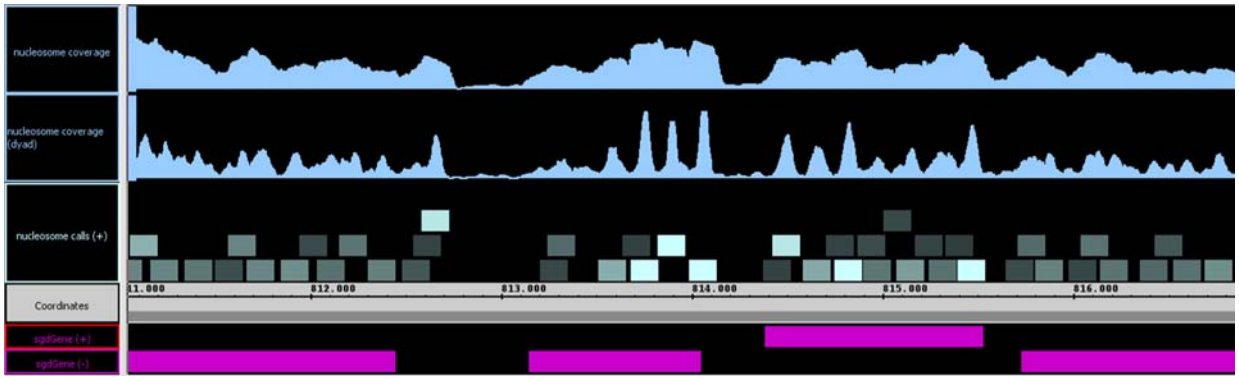


Figure A9. Nucleosome calling

Nucleosome calls over the reads coverage map of nucleosomal DNA, displayed in the chromosome 16 in the yeast genome. The top track shows raw paired-end reads coverage. Middle track shows the coverage map only taking into account the central 40 bp around the nucleosome dyad for each read. The different intensities of the blue-colored boxes account for different scores of the nucleosome calls (lighter blue indicates better positioned nucleosomes, darker blue indicates fuzzy or weak nucleosomes). Coordinates of the chromosome and genes are shown at the bottom.

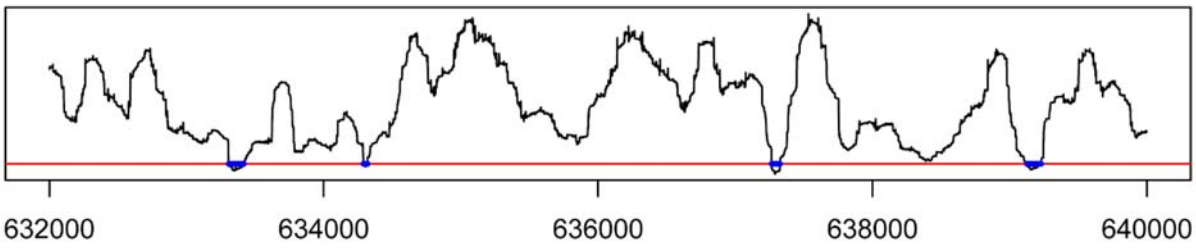


Figure A10. LRs detection in naked DNA

MNase digestion profile of naked DNA in a region of chromosome 16 (632,000–640,000). Horizontal red line marks the percentile 2.5 of the coverage and regions highlighted in blue correspond to the LRs identified within this threshold.

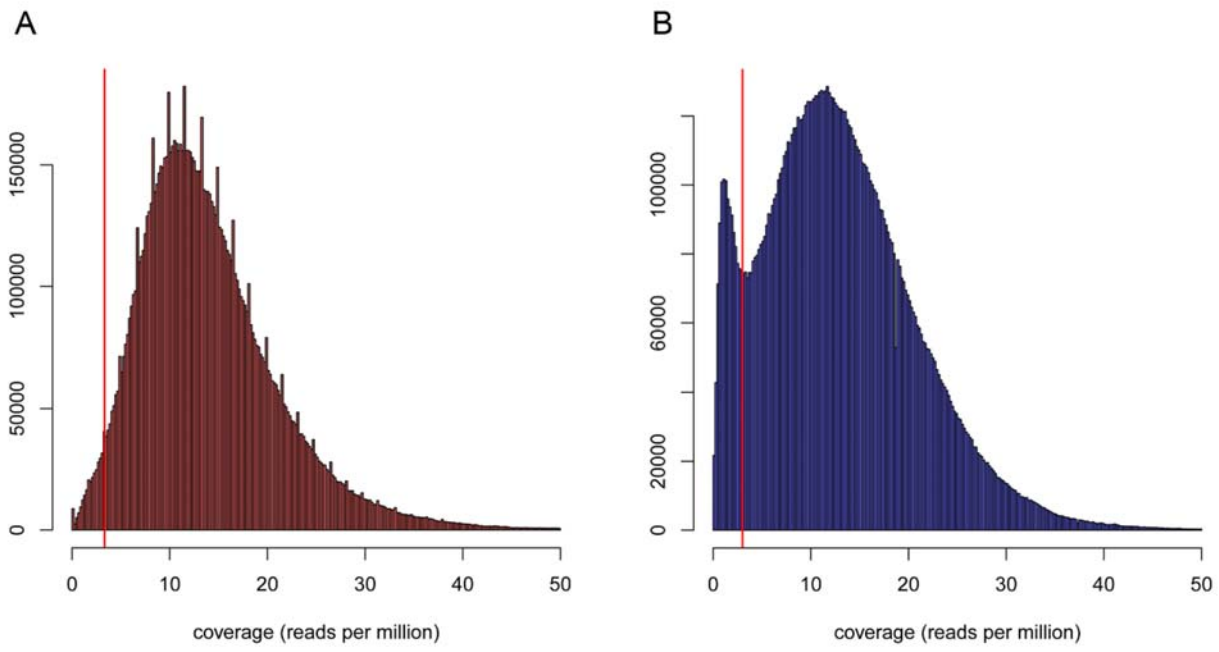


Figure A11. Coverage distribution

Histograms of reads coverage in naked DNA (left) and nucleosomal DNA (right). Y-axis represents the count number of a given coverage value detected on the genome. Percentile lines shown in red are 2.5% on left and 10% on the right.

Additional Tables

Table A1. Frequency of MNase non-preferred tetramers at the cutting sites

Naked DNA	ratio	p-val	Nucleosomal DNA	ratio	p-val (<)
AACT.AGTT	0.062	10^{-4}	AGGA.TCCT	0.062	4.00×10^{-4}
GGAA.TTCC	0.078	$< 10^{-18}$	AGCA.TGCT	0.062	2.00×10^{-4}
AGAT.ATCT	0.078	$< 10^{-18}$	ACCT.AGGT	0.065	1.51×10^{-3}
ACCA.TGGT	0.082	$< 10^{-18}$	AAGC.GCTT	0.097	6.00×10^{-4}
AAGT.ACTT	0.103	$< 10^{-18}$	TCCA.TGGA	0.098	10^{-4}
AGAA.TTCT	0.109	$< 10^{-18}$	ACCA.TGGT	0.121	9.00×10^{-4}
AAGA.TTCT	0.117	$< 10^{-18}$	AAGG.CCTT	0.196	1.11×10^{-3}
ATCA.TGAT	0.149	$< 10^{-18}$	AAGA.TCTT	0.211	$< 10^{-18}$
TGAA.TTCA	0.229	$< 10^{-18}$	AAGT.ACTT	0.227	1.21×10^{-3}
AAAA.TTTT	0.251	$< 10^{-18}$	AACA.TGTT	0.249	1.21×10^{-3}

Experimentally detected and expected frequency ratio of different MNase-non-preferred tetramers at the cutting sites in both naked (left) and nucleosomal (right) DNAs. Shown tetramers correspond to those less frequently observed than expected by a random model and they are not preferentially digested by MNase. The unfavorable cut sites have been selected considering the lowest ratio values between expected and random presence and statistically significant ($p < 10^{-4}$ for naked DNA and $p < 10^{-3}$ for nucleosomal DNA).

2. Impact of Methylation on the Physical Properties of DNA

Supplementary material for this work includes:

- Supplementary Methods
- Supplementary Figures

Supplementary Materials and Methods

Oligonucleotides

The oligonucleotides were synthesized and gel-purified by Sigma-Aldrich. As a positive control for multimerization-cyclization, we designed a 21bp sequence with 5'sticky ends derived from (1), which we called *PosSeq*: sense strand 5'-d(GAAAAAACGGGCGAAAAACGG) and antisense strand 5'-d(TCCCGTTTTTCGCCCGTTTTT) (here the underlined part of the sequence is complementary in both strands, with the overhanging bases at each end being complementary with the overhanging bases in the other strand, thus allowing for oligomerization). Conversely, as a negative control, we selected a oligo that was previously reported to have poor circularization properties (2), *NegSeq:* sense 5'-d(GCAAATATTGAAAAC) and antisense 5'-d(GCGTTTTCAATATTT).

The circularization favoring sequence was subsequently methylated at different positions, in a central cytosine on a CpG dinucleotide, *PosSeq_1met2*: sense 5'-d(GAAAAAACGGG**C**GAAAAACGG) and antisense 5'-d(TCCCGTTTTT**C**GCCCGTTTTT) (the C in bold-face corresponds to the methyl-cytosines). We also selected a set of molecules containing 2, 3 and 4 sequence repetitions of the 21bp *PosSeq* oligo (named *oligo 42*, *oligo 63* and *oligo 84*, respectively) as references for the size assignment.

Circularization Assays

Phosphorylation and radioactive labeling

Every single stranded oligonucleotide (1 nmol) was 5'-end-phosphorylated with 40U of T4 Polynucleotide kinase (New England Biolabs) by incubation at 37ºC in 1x T4 DNA Ligase Reaction Buffer. In case of radioactive phosphorylation, one picomole of DNA was mixed with 2.2 μmol of γ–32ATP (6 mCi/mL), 3U of T4 polynucleotide kinase and incubated for 1 hour at 37°C. The reaction mixture was purified using *MicroSpin G-25* columns (GE Healthcare) and resuspended in T4 DNA Ligase Reaction buffer.

Annealing and ligation

The complementary strands were denatured at 90 ºC and subsequently annealed by gradually decreasing the temperature to room conditions. Double stranded oligonucleotides were then multimerized with 400U of T4 DNA ligase (New England Biolabs) by an overnight incubation at room temperature. The reaction was stopped by inactivating the ligase at 65ºC for 10 minutes. The DNA was ethanol precipitated and dissolved in 10 μl of DNAse free water.

Digestion

The ligation samples were then digested with 20U of *Exonuclease III* (New England Biolabs) for 30 minutes at 37ºC. The reaction was stopped by inactivating the nuclease at 65ºC for 20 minutes.

Two-Dimensional Gel Electrophoresis

The ligation/digestion products were loaded on a 5% polyacrylamide native gel (19:1 acrylamide:bisacrylamide) and electrophoresed at room temperature at 4 V/cm in TBE buffer. After separating the different DNA species on the first dimension, the lanes of interest were excised from the gel and loaded on a second dimension gel (8% polyacrylamide) to separate circular and linear molecules of DNA. The separation of the two families of DNA was facilitated by the presence of chloroquine phosphate (250 μg/ml) that distorts the linear molecules of DNA in a different way from the circular ones (3). The gels were stained with *SyBr Safe* (Invitrogen) and visualized on the *ImageQuant* system (GE Healthcare) under UV light (Fig. S2).

Atomic Force Microscopy (AFM)

The ligation products were ethanol precipitated, resuspended in AFM Buffer (40mM HEPES, 10mM $MgCl_2$) and analyzed with AFM in tapping mode. AFM enabled us to have a rough estimate of the size of DNA molecules (see Fig. 2B) and confirmed the formation of minicircles in the experiments.

Linear and Circular DNA Size Determination

To determine the size of linear and circular DNA species more accurately, we performed individual circularization assays with *PosSeq*, *42*, *63* and *84* oligos, separately. Every sample was resolved in both 1D and 2D polyacrylamide gels (Fig. S3). By comparing the bands appearing or disappearing in the first and second dimension gels of the four samples, we were able to determine the exact size of every band. The actual size of the smallest circular molecules we obtained was determined to be 105 bp, with the following molecules increasing by 21 bp each.

Circular dichroism (CD) experiments

In order to rule out the hypothesis that massive cytosine methylation might induce a local change towards a Z-form of the DNA that could ultimately explain the functional role of CpG islands (4), we analyzed the secondary structure of the $d(CpG)_7$ oligonucleotide (with none or all the CpG steps methylated) by CD spectroscopy (Fig. S7). The DNA was exposed to an increasing ionic strength to force the B to Z transition. The transition B→Z was clearly favored in methylated DNA compared with the normal one, but in any case it occurred far beyond physiological conditions (Fig. S7), demonstrating that the transition to the Z-form is not the underlying principle for the physiological role of CpG methylation.

***In Vitro* Nucleosome Reconstitution**

Histone octamer purification

Histone octamers were purified from the nuclei of mammalian cells by a similar procedure described in detail elsewhere (5).

Accumulation of DNA substrate

The 146bp "601.2 sequence" (6, 7) was PCR amplified and cloned into the *PCR Script Amp+* vector (Stratagene), and subsequently used as a template for further amplification by PCR using *AccuPrime* Taq Polymerase (Invitrogen). Following PCR amplification, the oligo was agarose gel isolated and column purified (GE Healthcare). The DNA was ethanol precipitated, resuspended in water and quantified using *Nanodrop* spectrophotometer (Thermo Scientific)

In case of radioactive labeling, one picomole of DNA was mixed with 2.2 μmol of γ–32ATP (6 mCi/mL), 3U of T4 polynucleotide kinase and incubated for 1 hour at 37°C. The reaction mixture was purified using *MicroSpin G-25* columns (GE Healthcare) and resuspended in T4 DNA Ligase Reaction buffer.

CpG methylation of 601.2 sequence

The 601.2 sequence was methylated at all 12 CpG dinucleotides using the CpG methyltransferase MSssI (New England Biolabs). Briefly, 20μg of the 601 oligo was treated with 60U MSssI in the presence of 160 μM SAM (S-Adenosyl Methyionine) at 37ºC for 30 min. The reaction was replenished with another 60U MSssI and 160 μM SAM and incubated further for an hour at 37ºC. The reaction was stopped by incubation for 20 minutes at 65ºC. The CpG methylated oligo was purified from the enzyme using DNA purification columns (GE Healthcare). The localized methylation was verified by the sequencing of bisulfite treated DNA sample, which confirmed that all the cytosines in CpG dinucleotides were indeed methylated.

Full methylation of 601.2 sequence

Full cytosine methylation of the 601.2 sequence was achieved by replacing CTP with methyl-CTP (Fermentas) in the nucleotide mix during PCR amplification, using *Pfx* Taq Polymerase (Invitrogen). Subsequent agarose gel isolation, ethanol precipitation, and quantification were performed exactly as for the unmethylated 601 oligo.

Reconstitution reaction

Hydroxyapatite-purified histones were stored in aliquots in 2M NaCl/phosphate buffer pH 6.7 at -80ºC. Histones were thawed on ice and centrifuged 10,000 x g for 20 min at 4ºC to remove aggregates (improperly folded histones). Histones were subsequently quantified by measuring their absorption at 230nm.

The reconstitution reactions were prepared following a similar procedure as described in (8). All DNA and histones to be used in the reactions were freshly quantified immediately prior to use. Briefly, 750 ng of the respective forms of the 601.2 oligo (unmethylated control, CpG methylated, and full cytosine methylated)

were brought to 2M NaCl by adding an equal volume of 4M NaCl. The DNA was further mixed with histones at histone:DNA ratios ranging from 1.0-3.0 (w/w). The final volume of the reaction was brought to 35 µl using 2M NaCl / 50mM Tris pH 8.0 /1mM EDTA.

Each reconstitution reaction was mixed and transferred to a dialysis chamber (membrane 3,500 MWCO, Pierce). An initial volume of 100 ml of 2M NaCl /50mM Tris pH 8.0 / 1mM EDTA was diluted to 0.1M NaCl with continual addition of 50mM Tris pH 8.0 / 1mM EDTA to a final volume of 2 L using a peristaltic pump set at a flow rate of 40-60ml / hr at 4ºC. The dialyzed reaction was transferred to a microtube and stored at 4ºC.

Gel mobility shift assays

Nucleosome reconstitution was analyzed on 5% native polyacrylamide gels, which were pre-electrophoresed for 1hour at 100V at 4ºC in TBE. A 30% sucrose solution was added to the reconstitution reactions as a loading buffer immediately prior to loading the gel. The gels were run at 40 V for 6 hours at 4ºC and stained for 20 minutes in *SyBr Safe* (Invitrogen) diluted in 1x TBE. The intensities of the bands corresponding to nucleosome-bound or free DNA were subsequently estimated using the *ImageQuant* system (GE Healthcare) under UV light (see Fig. 4B).

In case of radioactive DNA, the bands were size-cut and the number of radiolabelled DNA molecules was then measured using a scintillation counter. In addition, the band intensities were also measured by densitometry using the *PhosphorImager* system (GE Healthcare) (see Fig. 4A).

Molecular Dynamics Simulations

MethylCytosine Parameters

Bonded parameters for ^{Me}C were based on the CHARMM27 parameters for ^{Me}C(9). Charges were derived from RESP fit at 6-31G** level starting from an optimized mC capped at the C1’ using the procedure described in reference (10). Library with parameters for d^{Me}C are available upon request.

Molecular Dynamics

DNA oligos were neutralized with Na+ counterions and hydrated with a octahedral box extending at least 10 Å in every direction from the DNA. The system was then optimized, thermalized (T = 298 K), and pre-equilibrated using a multistep protocol(11) in which we doubled the equilibration windows. The solvated systems were then allowed to evolve without restrains for an additional 100 ps before starting production runs.

Mesoscopic flexibility descriptors

Local parameters

MD simulations were projected into helical reference space to explore the sampled values of roll, twist, tilt, rise, shift and slide (12) for each of the ten unique base steps (plus methylated variants) in all unique tetramer environments. Equilibrium and stiffness parameters were obtained by following Lankas's procedure (13). Accordingly, MD trajectories were projected into a helical reference system to obtain equilibrium values and to derive the covariance matrix, which was then inverted following Einstein's equation to recover stiffness matrices, from which a mesoscopic estimate of the energy associated to a given deformation can be easily computed (see equation 1):

$$E = 0.5\sum_{i=1}^{6}\sum_{j=1}^{6} f_{ij}\Delta X_i \Delta X_j;\ \text{with the elements } f_{ij} \text{ corresponding to:}$$

$$\Theta = k_B T \subset^{-1} = \begin{pmatrix} k_w & k_{wr} & k_{wt} & k_{ws} & k_{wl} & k_{wf} \\ k_{wr} & k_r & k_{rt} & k_{rs} & k_{rl} & k_{rf} \\ k_{wt} & k_{rt} & k_t & k_{ts} & k_{tl} & k_{tf} \\ k_{ws} & k_{rs} & k_{ts} & k_s & k_{sl} & k_{sf} \\ k_{wl} & k_{rl} & k_{tl} & k_{sl} & k_l & k_{lf} \\ k_{wf} & k_{rf} & k_{tf} & k_{sf} & k_{lf} & k_f \end{pmatrix} \quad (1)$$

where k_b is the Boltzman constant, T is the absolute temperature, E is the energy associated to the deformation ΔX, and k stands for the different stiffness constants defining the 36 elements of the stiffness matrix (Θ) (twist (w), roll (r), tilt (t), rise

(s), slide (l) and shift (f)) at the dinucleotide level (in different tetramer environments) obtained by inversion of the MD-associated covariance matrix (⊂).

Global parameters

Global descriptors of equilibrium geometry and deformability were obtained for the different 18-mer oligonucleotides considering the 8mer central portion and Lavery's definitions (12) combined with Lankas's procedure (14). Descriptors considered here describe the ability of short tracts of DNA to bend in different directions (bending rigidity) and to undertwist/overtwist (twisting rigidity). They are calculated by considering fluctuations on this measures calculated at different oligomer lengths (up to a maximum of the full 18mer) following procedure described in (14), results shown in Fig. S1 are shown as the average for the twisting and bending rigidity of all possible 8mer combinations in each of the depicted 18mers.

Monte Carlo calculations of circularization efficiency

Theoretical J factors are calculated on the basis of the mesoscopic model described in methods (average and force constants for shift, slide, rise, tilt, roll and twist) and following a Monte Carlo procedure as described in the work of Olson and coworkers (15). Accordingly, J factors were calculated for PosSeq using different repeats of this sequence in the interval where circularization is successful experimentally (selected range of 126 to 210 bp). Since closed conformations are unlike to be sampled during the MC, efficiency was improved by: i.) Implementing Olson's gaussian sampling and ii.) using half-chain sampling enhancement. Both of them described in reference (15), labelling of an accepted conformation as closed was carried out again as described in reference (15). The J index is calculated as:

$$J = \frac{M_{closed}}{QM} \qquad (2)$$

where M_{closed} is the number of events in closed conformation, M is the number of conformations not closed and Q is defined in equation 11 of (15) as a normalization factor for the conditions imposed for determining a closed conformation. Note again

that since we are interested in relative rather than absolute J-factors arbitrary selections on when a conformation is “closed” do not affect our results.

Supplementary Figures

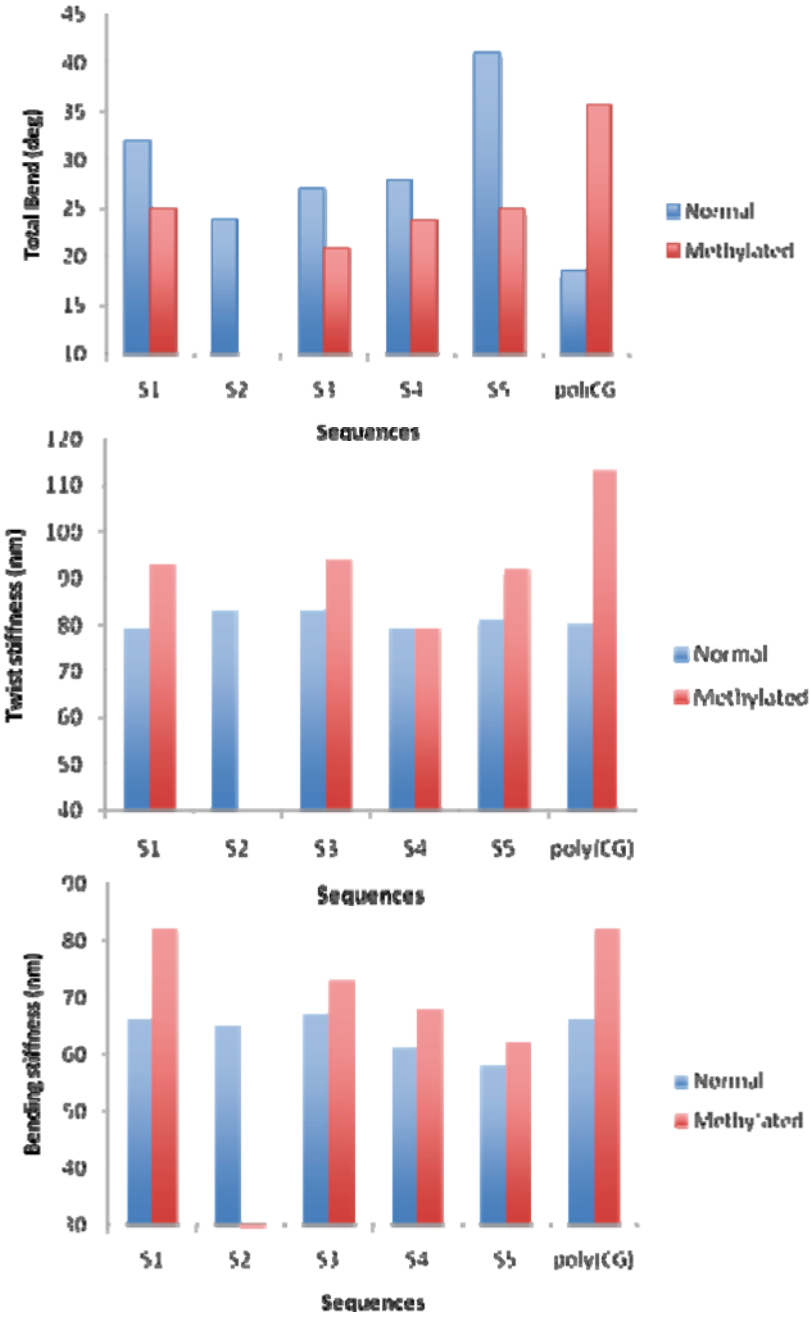


Figure S1. Description of the overall bend observed in 18mers in presence and absence of methylation (global bend as defined in (12)). Global twisting and bending rigidities are reported following work in (14). Results presented here show the rigidification of the sequence due to the presence of methylated d(CpG) steps. Note that sequence S2 (see Table S1 for definitions) has no d(CpG) steps.

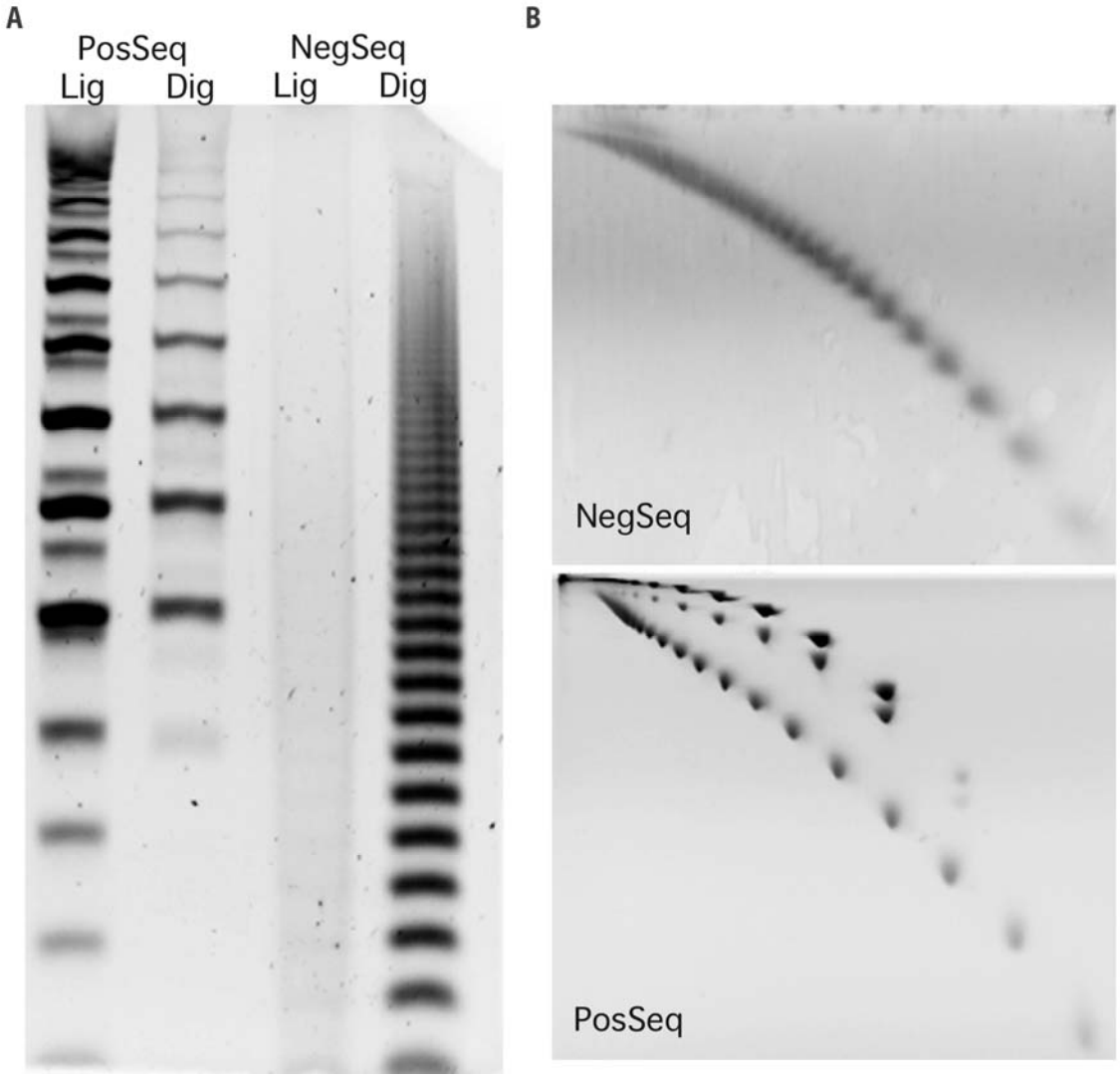


Figure S2. Circularization assays. (A) 5% polyacrylamide 1D-gel showing ligation (Lig) and digestion (Dig) products for the 21bp favoring (*PosSeq*) and 15bp non-favoring (*NegSeq*) cyclization oligonucleotides, respectively. The observed ladder patterns correspond to the different size multimers of the duplexes (both linear and circular). After digestion, only Exonuclease III-resistant DNA circles were observed. Thus, for the negative control sample, which is not expected to form circles, all the ligation products were digested. (B) 8% polyacrylamide 2D-native polyacrylamide gels showing the different migration of two species of DNA (linear and circular) from the ligation products. In the case of the positive control, linear DNA molecules are positioned on the lower diagonal, whereas closed and nicked circular DNA molecules in the upper diagonal. Conversely, for the negative control, only linear molecules are forming.

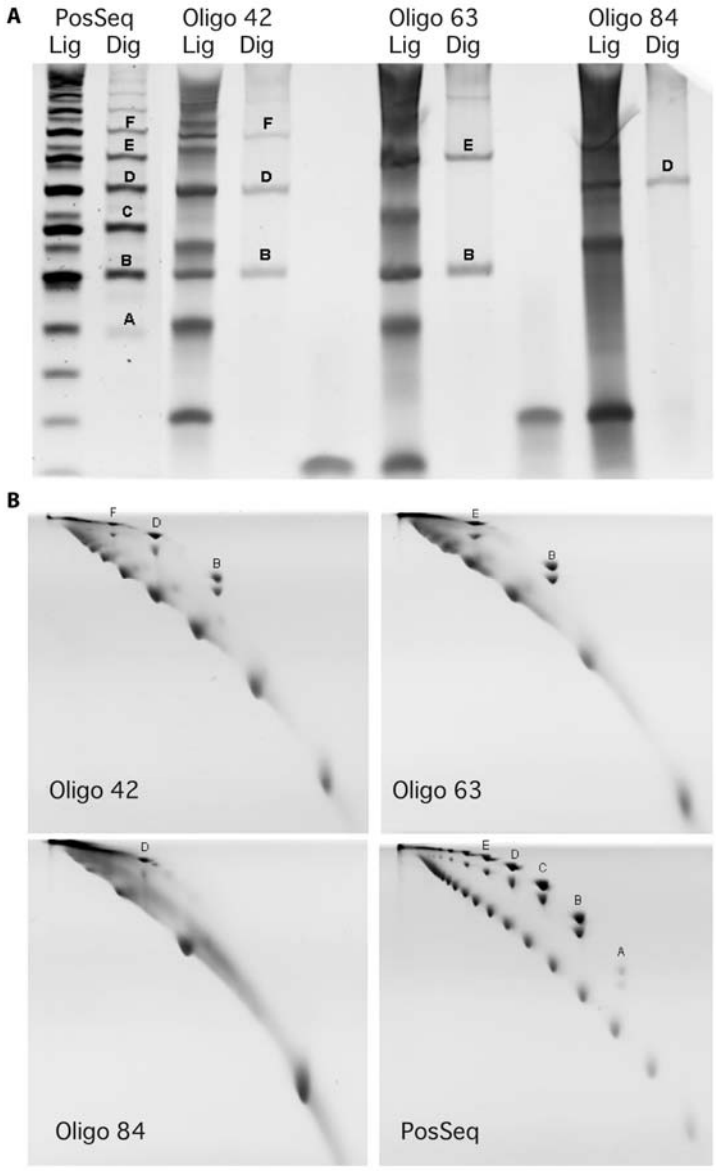


Figure S3. Circular and linear DNA size determination. (A) 1D gel showing ligation (Lig) and digestion (Dig) products for different size oligonucleotides. Lanes 1 and 2 contain *PosSeq* oligo ligation and digestion products, respectively; lanes 3, 6, 9: ligation products of oligo *42*, oligo *63* and oligo *84*, respectively; lanes 4, 7, 10: digestion products of the same oligos, respectively (corresponding to circular DNA); lanes 5, 8: duplexes of oligo *63* and oligo *84*. (B) 2D gels showing ligation products of oligo *42*, oligo *63*, oligo *84* and *PosSeq*, respectively. By comparing those bands appearing or disappearing on the different gels, we were able to calculate the exact size of each band.

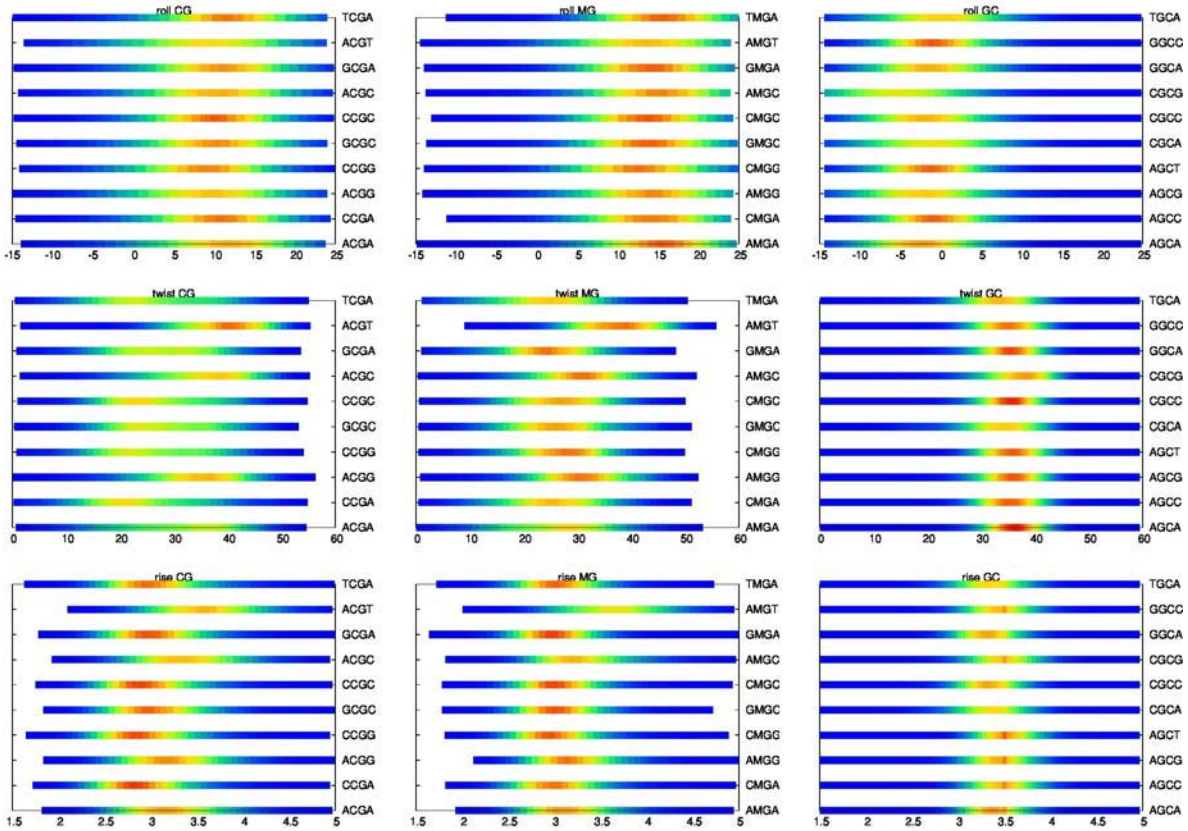


Figure S4. Histograms of roll(degree), twist(degree) and rise(Å) helical parameters for d(CpG), d(^{Me}CpG) and d(GpC). Each base pair step is described in all possible tetrad environments (denoted on the right hand axis 5'-ABCD-3'). Colours range from red (highly populated value) to blue (low populated).

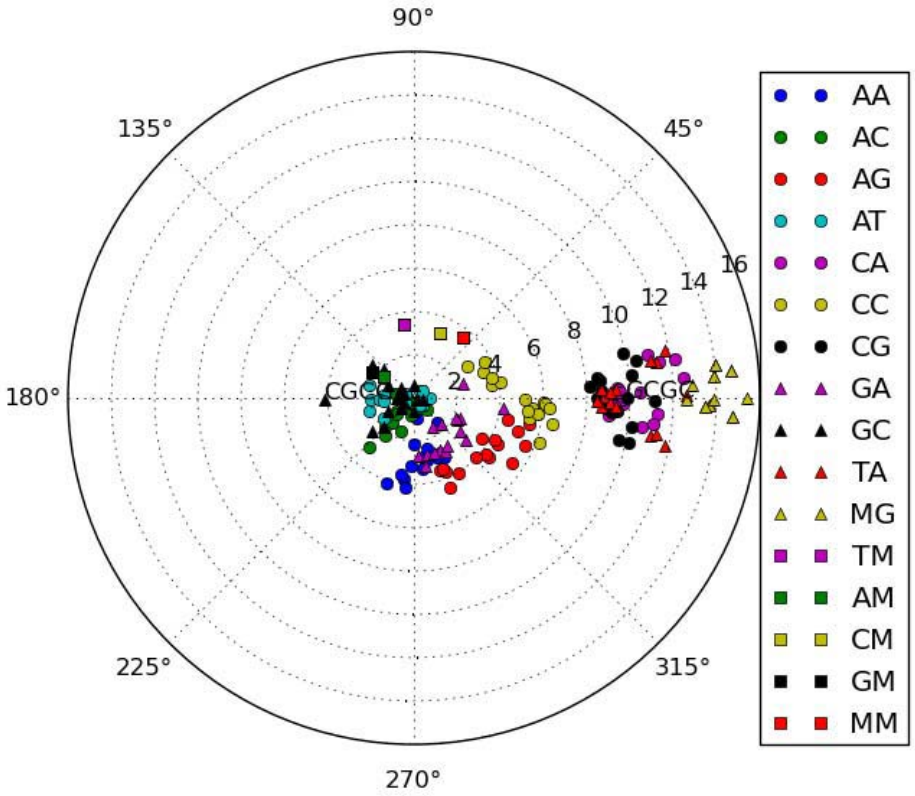


Figure S5. Bending direction and magnitude for the 10 base pair steps. Angles of ~0° denote bending towards the major groove and ~180° indicate bending towards the minor groove. For each base pair step, different data points represent the different nearest neighbour environments. A label has been added to d(GC) in d(CGCG) environment and d(CG) in d(GCGC) environment to note the counteracting effect leading to low overall curvature in poli(CpG) oligomers.

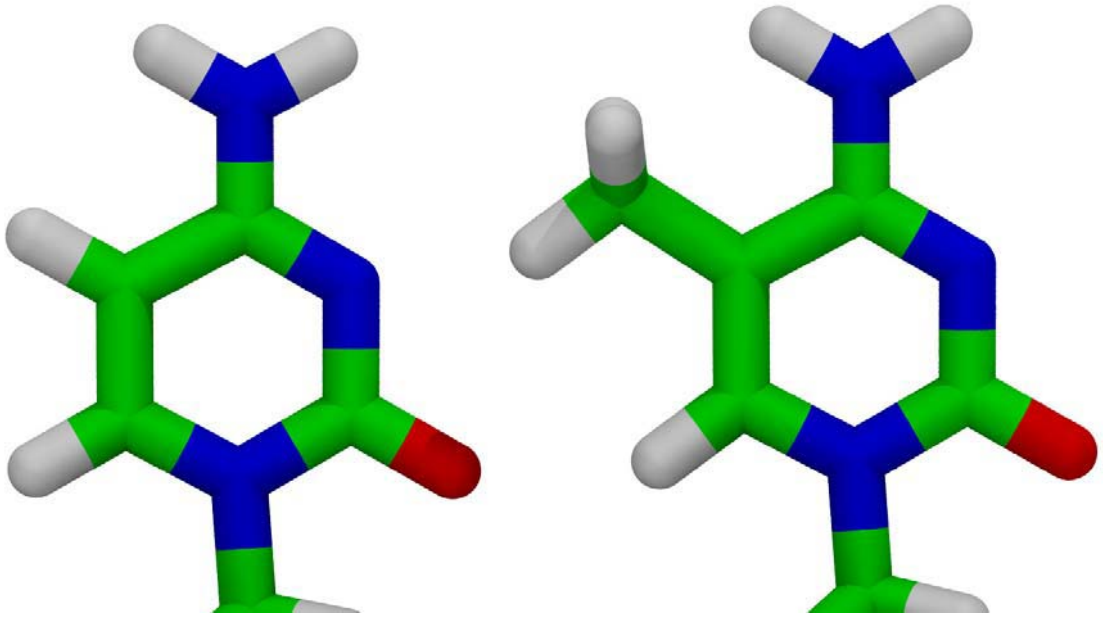


Figure S6. Chemical representation of Cytosine (left) and 5-methylCytosine (right). Color coding is: green for carbon atoms, red for oxygens, blue for nitrogens and white for hydrogens.

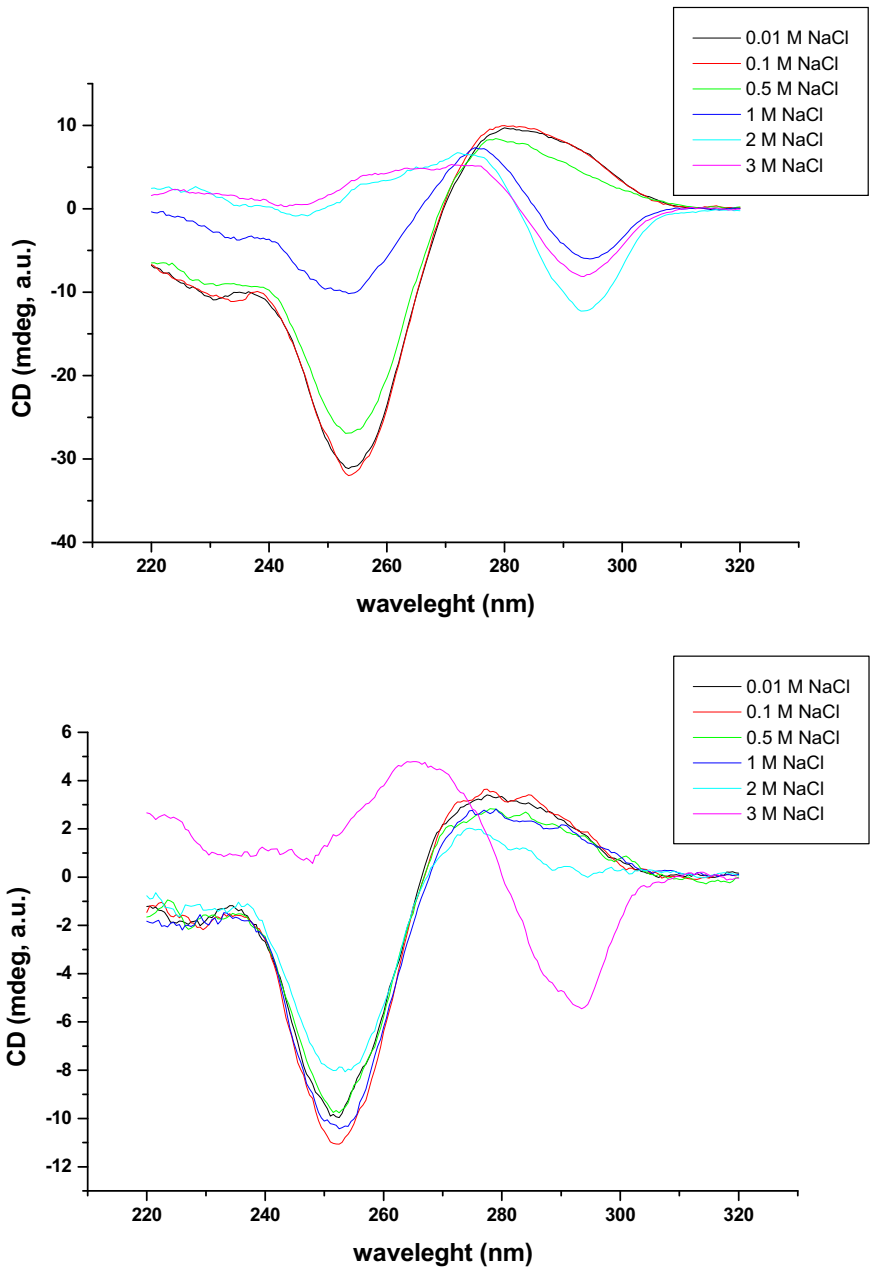


Figure S7. Circular Dichroism spectra monitoring the B to Z DNA transition for sequence d(^{Me}CpG)$_7$ (TOP) and d(CpG)$_7$ (BOTTOM): Curves are shown for different Ionic strength. The transition denoted by a shift in the peaks can be observed for concentrations much higher than those observed in physiological conditions, thus ruling out the change in conformation (B to Z) as the driving principle affecting the physical properties changes in d(CpG)n.

Supplementary Tables

Sequence name	Sequence identifier	Length	Sequence (5'→3')	Simulated time
Poli(MG)	P1/PM1	18mer	GMGMGMGMGMGMGMGMGM	100ns
Seq1*	S1	18mer	GCCTATAAA**MG**CCTATAA	100ns
Seq2*	S2	18mer	CTAGGTGGATGACTCATT	100ns
Seq3*	S3	18mer	CA**MG**GAAC**MG**GTTC**MG**TG	100ns
Seq4*	S4	18mer	GG**MGMG**CACCA**MGMGMG**G	100ns
Seq5	S5	18mer	GCCTATAAG**MGMG**TATAA	100ns
14MER-M1	M1	14mer	CGGA**MG**AC**MG**CGCG	100ns
14MER-M2	M2	14mer	CGGC**MG**AA**MG**CGCG	100ns
14MER-M3	M3	14mer	CGGA**MG**GG**MG**AGCG	100ns
14MER-M4	M4	14mer	CGGC**MG**GA**MG**TGCG	100ns
14MER-M5	M5	14mer	CGGG**MG**CT**MG**AGCG	100ns
ABC-dataset**	ABC	18mer	Multiple (39 different oligomers)	≥50ns

Table S1. Different sequences simulated for this study. The character M represents a methylated cytosine, all the corresponding non-methylated sequences were studied for the same length of time. The ABC dataset contains the most extensive set of trajectories and thus the most converged view on non-methylated sequences. Thus we considered its use for control of convergence in our dataset.
*These sequences correspond to the methylated counterparts of those published in reference 25 of the main text.
**These sequences correspond to the dataset generated in reference 20 of the main text.

Supplementary references

1. Podtelezhnikov, A. A., C. Mao, N. C. Seeman, and A. Vologodskii. 2000. Multimerization-cyclization of DNA fragments as a method of conformational analysis. Biophys J 79:2692-2704.
2. Payet, D., A. Hillisch, N. Lowe, S. Diekmann, and A. Travers. 1999. The recognition of distorted DNA structures by HMG-D: a footprinting and molecular modelling study. J Mol Biol 294:79-91.
3. Ulanovsky, L., M. Bodner, E. N. Trifonov, and M. Choder. 1986. Curved DNA: design, synthesis, and circularization. Proc Natl Acad Sci U S A 83:862-866.
4. Zacharias, W., A. Jaworski, and R. D. Wells. 1990. Cytosine methylation enhances Z-DNA formation in vivo. J Bacteriol 172:3278-3283.
5. Côté J., Utley R.T., and W. J.L.. 1995. Basic Analysis of transcription factor binding to nucleosomes. Methods Mol.Genet 6:108-129.
6. Lowary, P. T., and J. Widom. 1998. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. J Mol Biol 276:19-42.
7. Anderson, J. D., and J. Widom. 2000. Sequence and position-dependence of the equilibrium accessibility of nucleosomal DNA target sites. J Mol Biol 296:979-987.
8. Steger, D. J., T. Owen-Hughes, S. John, and J. L. Workman. 1997. Analysis of transcription factor-mediated remodeling of nucleosomal arrays in a purified system. Methods 12:276-285.
9. MacKerell, A. D., Jr., N. Banavali, and N. Foloppe. 2000. Development and current status of the CHARMM force field for nucleic acids. Biopolymers 56:257-265.
10. Cieplak, P., W. D. Cornell, C. Bayly, and P. A. Kollman. 1995. Application of the multimolecule and multiconformational RESP methodology to biopolymers: Charge derivation for DNA, RNA, and proteins. Journal of Computational Chemistry 16.
11. Shields, G. C., C. A. Laughton, and M. Orozco. 1997. Molecular dynamics simulations of the d(T center dot A center dot T) triple helix. Journal of the American Chemical Society 119:7463-7469.
12. Lavery, R., M. Moakher, J. H. Maddocks, D. Petkeviciute, and K. Zakrzewska. 2009. Conformational analysis of nucleic acids revisited: Curves+. Nucleic Acids Res 37:5917-5929.
13. Lankas, F., J. Sponer, J. Langowski, and T. E. Cheatham, 3rd. 2003. DNA basepair step deformability inferred from molecular dynamics simulations. Biophys J 85:2872-2883.
14. Lankas, F., J. Sponer, P. Hobza, and J. Langowski. 2000. Sequence-dependent elastic properties of DNA. J Mol Biol 299:695-709.
15. Czapla, L., D. Swigon, and W. K. Olson. 2006. Sequence-dependent effects in the cyclization of short DNA. J Chem Theory Comput 2:685-695.

3. Unraveling the hidden DNA structural/physical code provides novel insights on promoter location

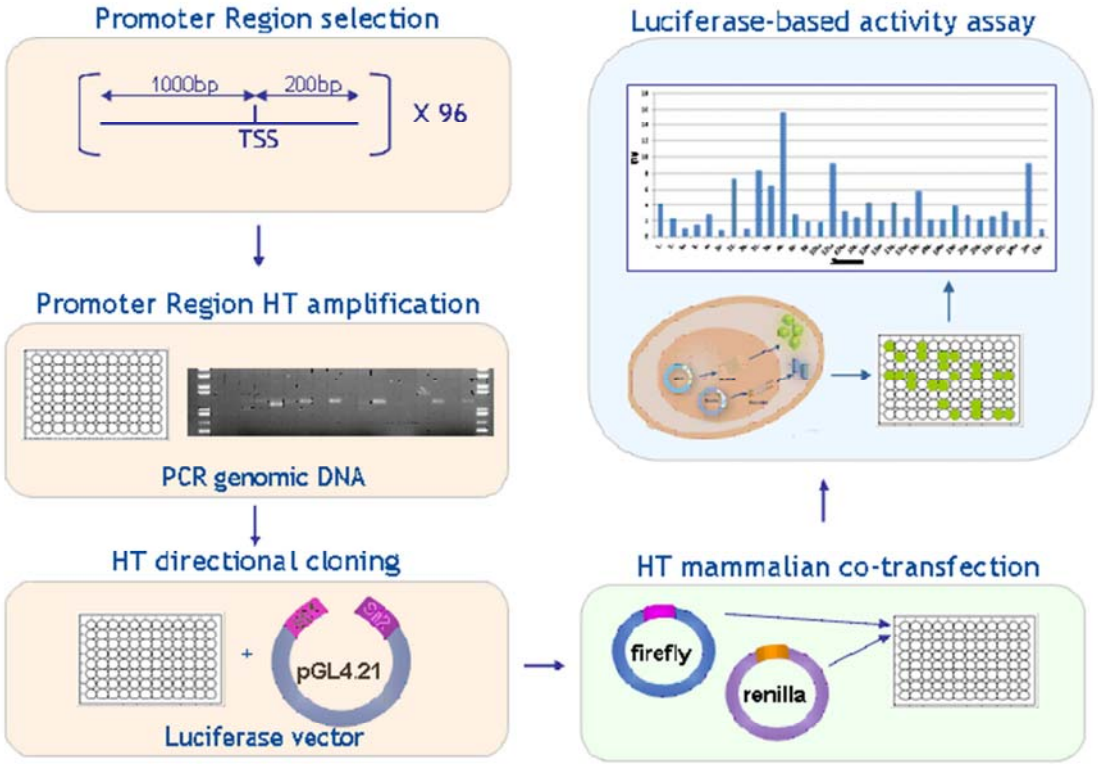
Supplementary material for this work includes:

- Supplementary Figure 1 and 2
- Supplementary Table 2

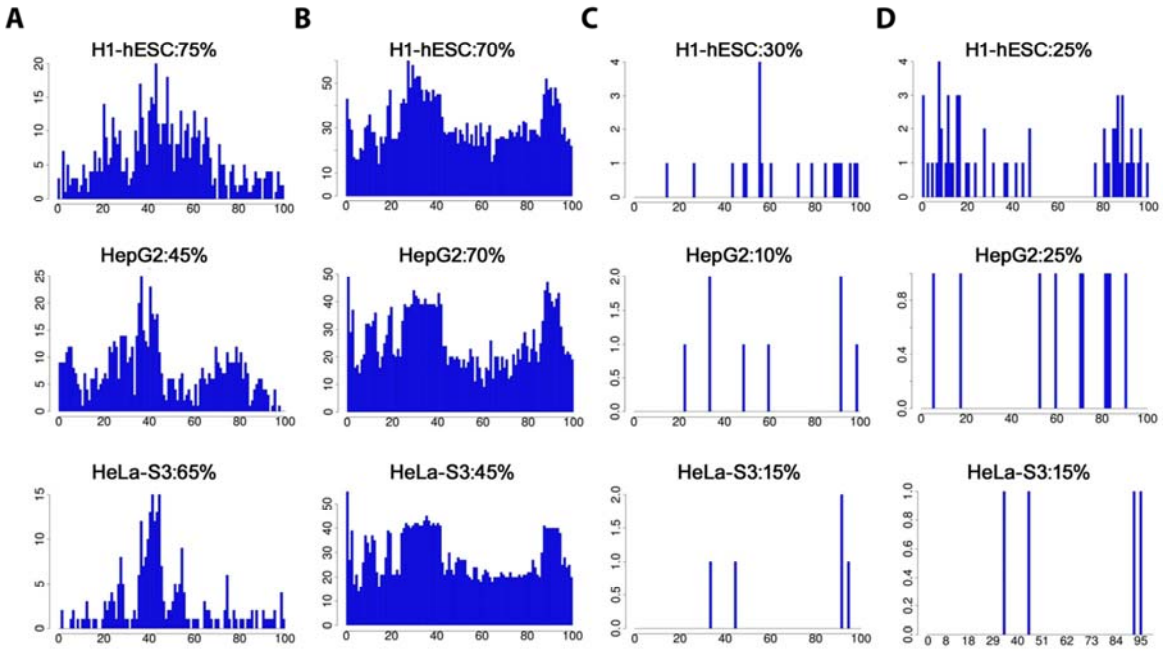
Additional materials are available in the editor's website http://dx.doi.org/10.1093/nar/gkt511

Unraveling the hidden DNA structural/physical code provides novel insights on promoter location

Elisa Durán, Sarah Djebali, Santi González, Oscar Flores, Josep Maria Mercader, Roderic Guigó, David Torrents, Montserrat Soler-López and Modesto Orozco



Supplementary Figure 1. Experimental approach for luciferase activity assays in a high-throughput approach. Five major steps: (1) selection and amplification or putative promoter regions (2) high-throughput (HT) PCR amplification of selected fragments from human genomic DNA (3) HT fragment cloning into the promoter-less vector upstream of the firefly luciferase encoding gene (4) transient co-transfection in mammalian cells with a constitutive luciferase expressing vector (renilla) for value normalization (5) firefly luciferase measurements to evaluate the potential promoter activity of selected fragments.



Supplementary Figure S2. Orthogonal support of predicted TSSs: RNA-seq analysis. Distribution of distinct sequenced tags from several representative RNA-seq experiments in H1-hESC, HepG2 and HeLa-S3 cell types based on cytosolic polyA+ transcripts. For every distinct most 5'-end of RNA-seq tag detected within and on the same strand as a particular promoter region, we increased the RNA-seq frequency of the percent distance bin corresponding to the distance between the RNA-seq tag 5'-end and the promoter region 5'-end. As the predicted promoter regions were 2000 bp long, each % distance bin includes 20 bp and thereby the TSS is expected to be located on the 50th distance bin (i.e. at 1000 bp from the region 5'-end). (a) PS+L+ subset 1. For most of the cell types, the major peak appears around the 40^{th} bin (i.e. 800 bp), closely matching with the prediction (b) PS+L- subset 2. We observe undefined peaks around the 10^{th}-20^{th} bins (200-400 bp). On the other hand, the number of RNA-seq tags is significantly higher than for subset 1 (c) PS-L+ subset 3 (d) PS-L- subset 4. ProStar negative PS- subsets clearly show an almost inexistent RNA-seq signal.

Unraveling the hidden DNA structural/physical code provides novel insights on promoter location

Elisa Durán, Sarah Djebali, Santi González, Oscar Flores, Josep Maria Mercader, Roderic Guigó, David Torrents, Montserrat Soler-López and Modesto Orozco

Supplementary Table 2. Correlation of physical deformability patterns with the CG content. ProStar positive predictions analyzed by CAGE that are located at ≥ 3500 bp distance to any annotated CpG island, based on UCSC Genome Browser. The relative distance to the closest annotated CpG islands is indicated as base pairs (bp). (*) The negative symbol refers to the strand orientation.

Subset	Chromosome	Predicted TSS location (GRCh37/hg19)	Tested region range 1,2kb (GRCh37/hg19)	Distance (bp) to the closest annotated CGI*
G1 (PS+/L+)	7	61.820.762	61820562..61821762	693.753
	1	40.335.364	40334364..40335564	13.577
	4	62.068.131	62067931..62069131	314.882
	5	138.897.705	138897505..138898705	42.879
G2 (PS+/L-)	4	147.576.411	147576211..147577411	-14.510
	14	60.981.788	60981588..60982788	-3.608
	18	76.752.776	76752576..76753776	-11.532
	19	920.551	920351..921551	4.237

4. Fuzziness and noise in nucleosomal architecture

Supplementary material for this work includes:

- Supplementary Tables
- Supplementary Figures

FUZZINESS AND NOISE IN NUCLEOSOMAL ARCHITECTURE

Oscar Flores, Özgen Deniz, Montserrat Soler-López and Modesto Orozco

SUPPLEMENTARY TABLES

default	**score_w**	**0.6**		**score_h**	**0.4**		
	FcF	FoF	WcW	WoW	M-F	M--W	+1_missing
R1 (2x)	226	89	1808	1674	106	605	42
R2 (2x)	183	59	1942	1606	86	464	20
As (2x)	277	167	1624	1419	171	658	34
Ov (2x)	204	205	1405	1285	314	840	56
Un (2x)	478	358	763	623	211	361	15
var1	**score_w**	**0.5**		**score_h**	**0.5**		
	FcF	FoF	WcW	WoW	M-F	M--W	+1_missing
R1 (2x)	149	63	1785	1618	103	608	42
R2 (2x)	109	46	1916	1574	74	476	20
As (2x)	136	113	1722	1476	133	696	34
Ov (2x)	211	210	1261	1147	345	809	56
Un (2x)	360	273	853	677	174	398	15
var2	**score_w**	**0.4**		**score_h**	**0.6**		
	FcF	FoF	WcW	WoW	M-F	M--W	+1_missing
R1 (2x)	287	152	1433	1362	149	562	42
R2 (2x)	187	79	1643	1415	104	446	20
As (2x)	238	185	1410	1259	194	635	34
Ov (2x)	341	354	1016	935	424	730	56
Un (2x)	431	335	726	596	198	374	15

Supplementary Table S1. Nucleosomes call evaluation based on nucleR TSS clustering by applying different thresholds. The number of genes clustered in the main nucleosome pattern annotations are displayed according to the default parameters score_w = 0.6 and score_h=0.4; score_w=0.5 and score_h=0.5 (*var1*); and score_w=0.4 and score_h=0.6 (*var2*). Key: **R1 (2x)** – replica 1 paired-end dataset; **R2** – replica 2; **As** – asynchronized; **Ov** – MNase-overdigested; **Un** – underdigested.

	-1 Nucleosome			NFR		+1 Nucleosome	
	M	**F**	**W**	**Closed**	**Open**	**F**	**W**
R1 (1x)	648 9.68%	2544 38.02%	2992 44.71%	2858 42.71%	2207 32.98%	2405 35.94%	3779 56.47%
R2 (1x)	239 3.57%	2094 31.29%	395 59.06%	3726 55.68%	1860 27.79%	1390 20.77%	4895 73.15%
R1 (2x)	711 10.62%	1508 22.53%	4021 60.09%	2939 43.92%	2300 34.37%	1029 15.38%	5211 77.87%
R2 (2x)	550 8.22%	1070 15.99%	3852 57.56%	2742 40.97%	2040 30.48%	684 10.22%	4788 71.55%
As. (2x)	829 12.39%	1733 25.9%	3656 54.63%	2794 41.75%	2225 33.25%	1302 19.46%	4916 73.46%
Ov. (2x)	1154 17.24%	1886 28.18%	3235 48.34%	2559 38.24%	2328 34.79%	1321 19.74%	4954 74.03%
Un. (2x)	572 8.55%	2530 37.81%	2051 30.65%	2302 34.4%	1752 26.18%	1846 27.59%	3307 49.42%

Supplementary Table S2. Distribution of -1/+1 nucleosomes and NFRs classification for each sample. -1 nucleosomes are classified as missing (M), fuzzy (F) or well-positioned (W) and +1 nucleosomes are either F or W. NFRs can have open or closed configuration depending on the NFR width. Key: **1x** – Single End, **2x** – Paired End, **R1** – Replica 1, **R2** – Replica2, **As** – Asynchronous, **Ov** – Overdigested, **Un** – Underdigested.

	Same classification					Variable classification				
vs.	**Coverage**	**Cluster**	**-1 Nuc.**	**+1 Nuc.**	**NFR**	**Coverage**	**Cluster**	**-1 Nuc.**	**+1 Nuc.**	**NFR**
R1 (1x) R1 (2x)	6311 (94.31%)	2552 (38.14%)	4109 (61.4%)	4261 (63.67%)	4599 (68.72%)	23 (0.34%)	1513 (22.61%)	489 (7.31%)	448 (6.69%)	390 (5.83%)
R2 (1x) R2 (2x)	5969 (89.2%)	2399 (35.85%)	3480 (52%)	4301 (64.27%)	3751 (56.05%)	63 (0.94%)	1514 (22.62%)	791 (11.82%)	335 (5.01%)	591 (8.83%)
R1 (1x) R2 (1x)	5418 (80.96%)	2014 (30.1%)	3606 (53.89%)	4205 (62.84%)	4091 (61.13%)	222 (3.32%)	1994 (29.8%)	1208 (18.05%)	716 (10.7%)	710 (10.61%)
R1 (2x) R2 (2x)	5980 (89.36%)	3181 (47.53%)	4068 (60.79%)	4529 (67.68%)	4260 (63.66%)	93 (1.39%)	992 (14.82%)	433 (6.47%)	470 (7.02%)	231 (3.45%)
R1 (2x) As (2x)	5123 (76.55%)	2964 (44.29%)	4131 (61.73%)	4758 (71.1%)	4482 (66.98%)	228 (3.41%)	1563 (23.36%)	858 (12.82%)	819 (12.24%)	322 (4.81%)
R2 (2x) As (2x)	5781 (86.39%)	2927 (43.74%)	3874 (57.89%)	4393 (65.65%)	4157 (62.12%)	120 (1.79%)	1144 (17.1%)	616 (9.21%)	611 (9.13%)	212 (3.17%)
R1 (2x) Ov (2x)	5616 (83.92%)	2881 (43.05%)	3966 (59.26%)	4808 (71.85%)	4452 (66.53%)	91 (1.36%)	1649 (24.64%)	840 (12.55%)	857 (12.81%)	322 (4.81%)
R2 (2x) Ov (2x)	5874 (87.78%)	2754 (41.15%)	3612 (53.97%)	4443 (66.39%)	4155 (62.09%)	106 (1.58%)	1181 (17.65%)	680 (10.16%)	670 (10.01%)	217 (3.24%)
R1 (2x) Un (2x)	4177 (62.42%)	1377 (20.58%)	2597 (38.81%)	3426 (51.2%)	2917 (43.59%)	405 (6.05%)	2111 (31.55%)	1179 (17.62%)	1090 (16.29%)	705 (10.53%)
R2 (2x) Un (2x)	4971 (74.28%)	1355 (20.25%)	2376 (35.51%)	3202 (47.85%)	2767 (41.35%)	221 (3.3%)	1820 (27.2%)	1184 (17.69%)	1047 (15.65%)	586 (8.76%)
Ov (2x) Un (2x)	6342 (94.77%)	1915 (28.62%)	3105 (46.4%)	3737 (55.84%)	3279 (49%)	19 (0.28%)	1661 (24.82%)	721 (10.77%)	731 (10.92%)	547 (8.17%)
As (2x) Ov (2x)	6421 (95.95%)	3315 (49.54%)	4381 (65.47%)	5033 (75.21%)	4776 (71.37%)	14 (0.21%)	1229 (18.37%)	414 (6.19%)	484 (7.23%)	259 (3.87%)
As (2x) Un (2x)	5890 (88.02%)	1703 (25.45%)	2876 (42.98%)	3648 (54.51%)	3065 (45.8%)	79 (1.18%)	1908 (28.51%)	982 (14.67%)	882 (13.18%)	620 (9.26%)

Supplementary Table S3. Different pair-wise metrics of nucleosome similarity/dissimilarity. In order to obtain robust estimations of the similarity/dissimilarity of gene architectures between samples we defined the following metrics. *Coverage*: A gene is considered as stable if Pearson's correlation between two samples in the window -300:300 from the TSS is greater than 0.7; is considered as variable if correlation is smaller than 0.5. *Cluster*: We consider a gene stable if the cluster stays the same; we considered a significant variable architecture when 2 of the clustering dimensions (-1/NFR/+1) vary between samples. *+1/-1 Nucleosome*: We consider a nucleosome in the same classification if nucleR's classification is the same for two samples; we considered a gene variable if the absolute difference in nucleR's score is bigger than 0.25 points. *NFR*: we consider a gene stable if the classification of the NFR is the same (open/close/overlap/missing); we considered a change as significant if the change in distance between -1/+1 nucleosomes is more than 100bp. Genes which do not satisfy any of the two criteria are in considered out of the stability/variability threshold. Percentages are relative to the total number of genes in the SacCer3 genome. Key: **1x** – Single End, **2x** – Paired End, **R1** – Replica 1, **R2** – Replica2, **As** – Asynchronous, **Ov** – Overdigested, **Un** – Underdigested.

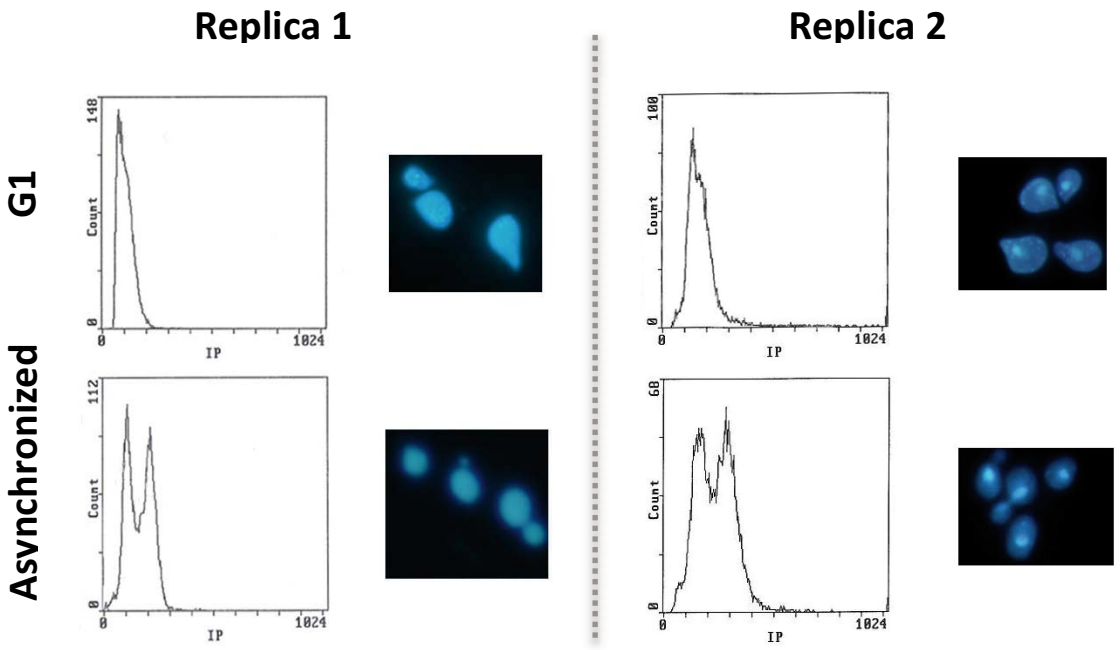
	Stable classification			Variable classification		
vs.	**Coverage**	**Δcalls**	**Δscore**	**Coverage**	**Δcalls**	**Δscore**
R1 (1x) R1 (2x)	5688 (85%)	3077 (45.98%)	6010 (89.81%)	169 (2.53%)	446 (6.66%)	33 (0.49%)
R2 (1x) R2 (2x)	5436 (81.23%)	3142 (46.95%)	5862 (87.6%)	194 (2.9%)	679 (10.15%)	39 (0.58%)
R1 (1x) R2 (1x)	3776 (56.43%)	2976 (44.47%)	5634 (84.19%)	486 (7.26%)	386 (5.77%)	59 (0.88%)
R1 (2x) R2 (2x)	5212 (77.88%)	3261 (48.73%)	5759 (86.06%)	341 (5.1%)	577 (8.62%)	53 (0.79%)
R1 (2x) As (2x)	3190 (47.67%)	2901 (43.35%)	4958 (74.09%)	869 (12.99%)	461 (6.89%)	96 (1.43%)
R2 (2x) As (2x)	4578 (68.41%)	2971 (44.4%)	5461 (81.6%)	459 (6.86%)	537 (8.02%)	72 (1.08%)
R1 (2x) Ov (2x)	4297 (64.21%)	3096 (46.26%)	5129 (76.64%)	463 (6.92%)	417 (6.23%)	114 (1.7%)
R2 (2x) Ov (2x)	5047 (75.42%)	3448 (51.52%)	5462 (81.62%)	427 (6.38%)	321 (4.8%)	91 (1.36%)
R1 (2x) Un (2x)	2550 (38.11%)	2493 (37.25%)	3938 (58.85%)	1176 (17.57%)	769 (11.49%)	262 (3.92%)
R2 (2x) Un (2x)	3678 (54.96%)	2647 (39.55%)	4279 (63.94%)	767 (11.46%)	508 (7.59%)	224 (3.35%)
Ov (2x) Un (2x)	6058 (90.53%)	2786 (41.63%)	5631 (84.15%)	121 (1.81%)	459 (6.86%)	48 (0.72%)
As (2x) Ov (2x)	6101 (91.17%)	3014 (45.04%)	6094 (91.06%)	91 (1.36%)	554 (8.28%)	17 (0.25%)
As (2x) Un (2x)	5280 (78.9%)	2462 (36.79%)	5294 (79.11%)	237 (3.54%)	828 (12.37%)	101 (1.51%)

Supplementary Table S4. Different pair-wise metrics of nucleosome similarity/dissimilarity in gene body regions. *Coverage*: a particular gene is considered 'stable' when Pearson's correlation between the whole gene body of two samples is greater than 0.7; otherwise, it is considered 'variable' when correlation is smaller than 0.5. Δ*calls*: a gene is considered 'stable' when it has the same number of nucleosome calls inside the gene body; otherwise, it is considered 'variable' when the difference of nucleosome calls between samples is more than 2 nucleosomes. Δ*score*: a particular gene is considered to be the same between two samples when the equivalent nucleosome calls have a mean difference of nucleR score lower than 0.15; otherwise, the gene is assigned as variable when this difference is higher than 0.25. Genes which do not satisfy any of the two criteria are in considered out of the stability/variability threshold. Percentages are relative to the total number of genes in the SacCer3 genome. *Note: the present metrics differ from those presented in the previous Supplementary Table 3, in order to account for an arbitrary number of nucleosome calls in gene body regions.* Key: **1x** – Single End, **2x** – Paired End, **R1** – Replica 1, **R2** – Replica2, **As** – Asynchronous, **Ov** – Overdigested, **Un** – Underdigested.

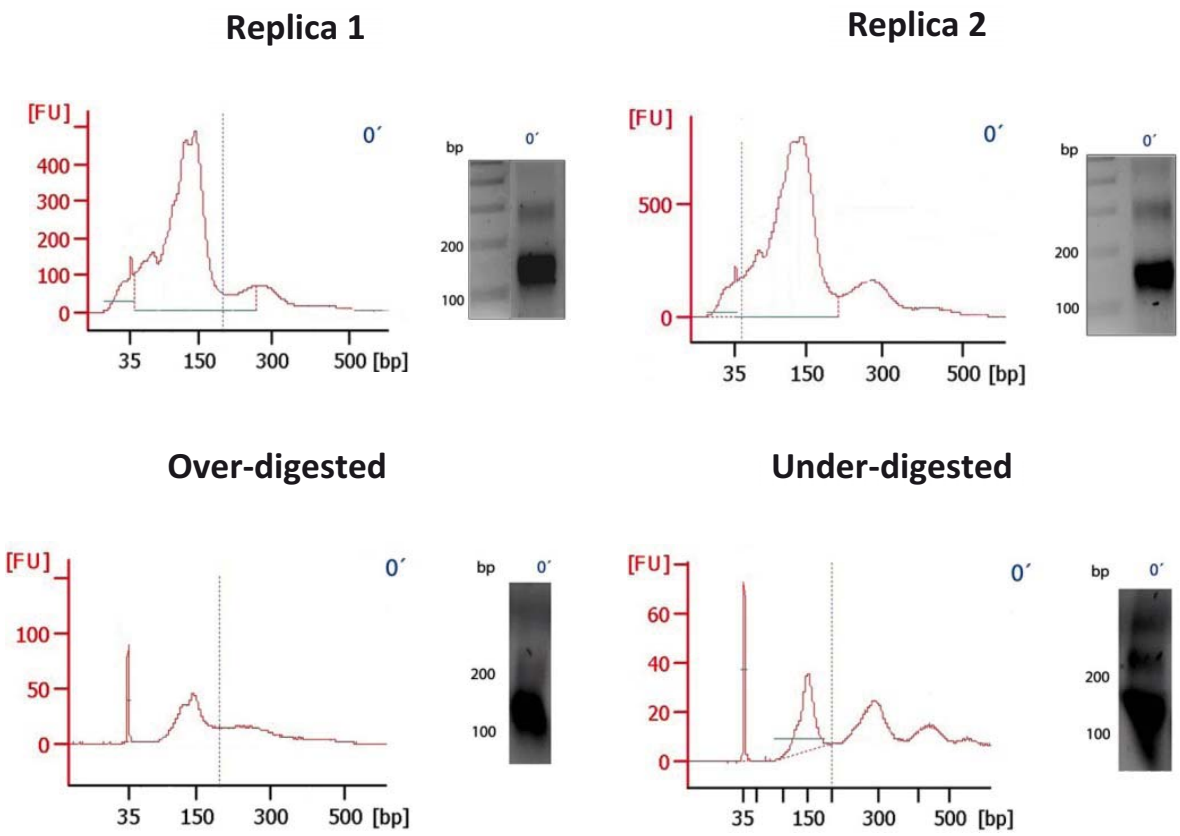
	Stable classification			Variable classification		
vs.	**Coverage**	**Δcalls**	**Δscore**	**Coverage**	**Δcalls**	**Δscore**
R1 (1x) R1 (2x)	6210 (92.80%)	3965 (59.25%)	5868 (87.69%)	81 (1.21%)	227 (3.39%)	27 (0.4%)
R2 (1x) R2 (2x)	5944 (88.82%)	2893 (43.23%)	5729 (85.61%)	118 (1.76%)	830 (12.4%)	50 (0.75%)
R1 (1x) R2 (1x)	5455 (81.52%)	3411 (50.97%)	5220 (78%)	273 (4.08%)	82 (1.23%)	92 (1.37%)
R1 (2x) R2 (2x)	5929 (88.60%)	4397 (65.71%)	5614 (83.89%)	174 (2.6%)	281 (4.2%)	76 (1.14%)
R1 (2x) As (2x)	5136 (76.75%)	3907 (58.38%)	4825 (72.10%)	364 (5.44%)	54 (0.81%)	176 (2.63%)
R2 (2x) As (2x)	5738 (85.74%)	4303 (64.30%)	5366 (80.19%)	242 (3.62%)	92 (1.37%)	108 (1.61%)
R1 (2x) Ov (2x)	5562 (83.11%)	3885 (58.05%)	4787 (71.53%)	209 (3.12%)	252 (3.77%)	179 (2.67%)
R2 (2x) Ov (2x)	5820 (86.97%)	4477 (66.90%)	5109 (76.34%)	223 (3.33%)	87 (1.3%)	118 (1.76%)
R1 (2x) Un (2x)	4344 (64.91%)	2890 (43.19%)	3928 (58.70%)	649 (9.7%)	179 (2.67%)	447 (6.68%)
R2 (2x) Un (2x)	4975 (74.34%)	3000 (44.83%)	3880 (57.98%)	445 (6.65%)	160 (2.39%)	453 (6.77%)
Ov (2x) Un (2x)	6248 (93.37%)	3248 (48.54%)	5254 (78.51%)	67 (1%)	118 (1.76%)	87 (1.3%)
As (2x) Ov (2x)	6353 (94.93%)	4059 (60.65%)	5982 (89.39%)	43 (0.64%)	290 (4.33%)	33 (0.49%)
As (2x) Un (2x)	5837 (87.22%)	3009 (44.96%)	4858 (72.59%)	163 (2.44%)	149 (2.23%)	170 (2.54%)

Supplementary Table S5. Different pair-wise metrics of nucleosome similarity/dissimilarity around TSSs (window -300:+300). Metrics used here are the same than in Supplementary Table S4.

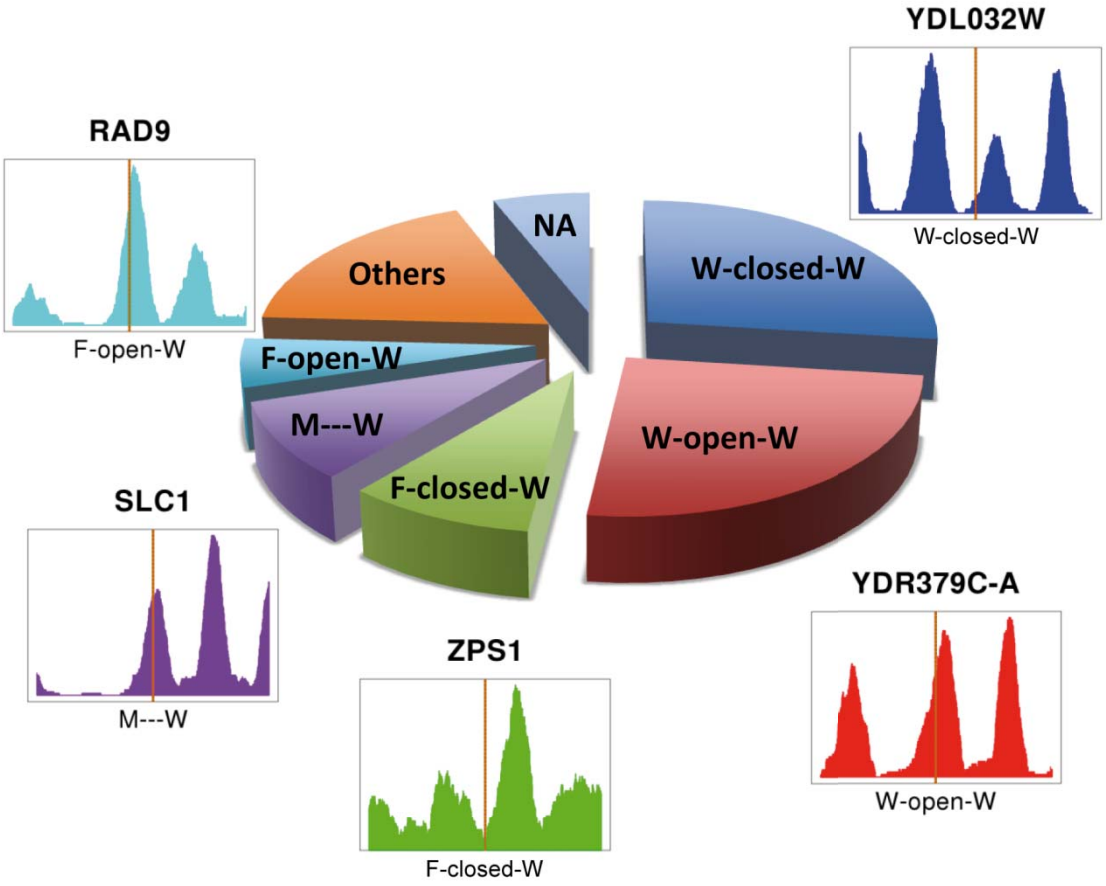
SUPPLEMENTARY FIGURES



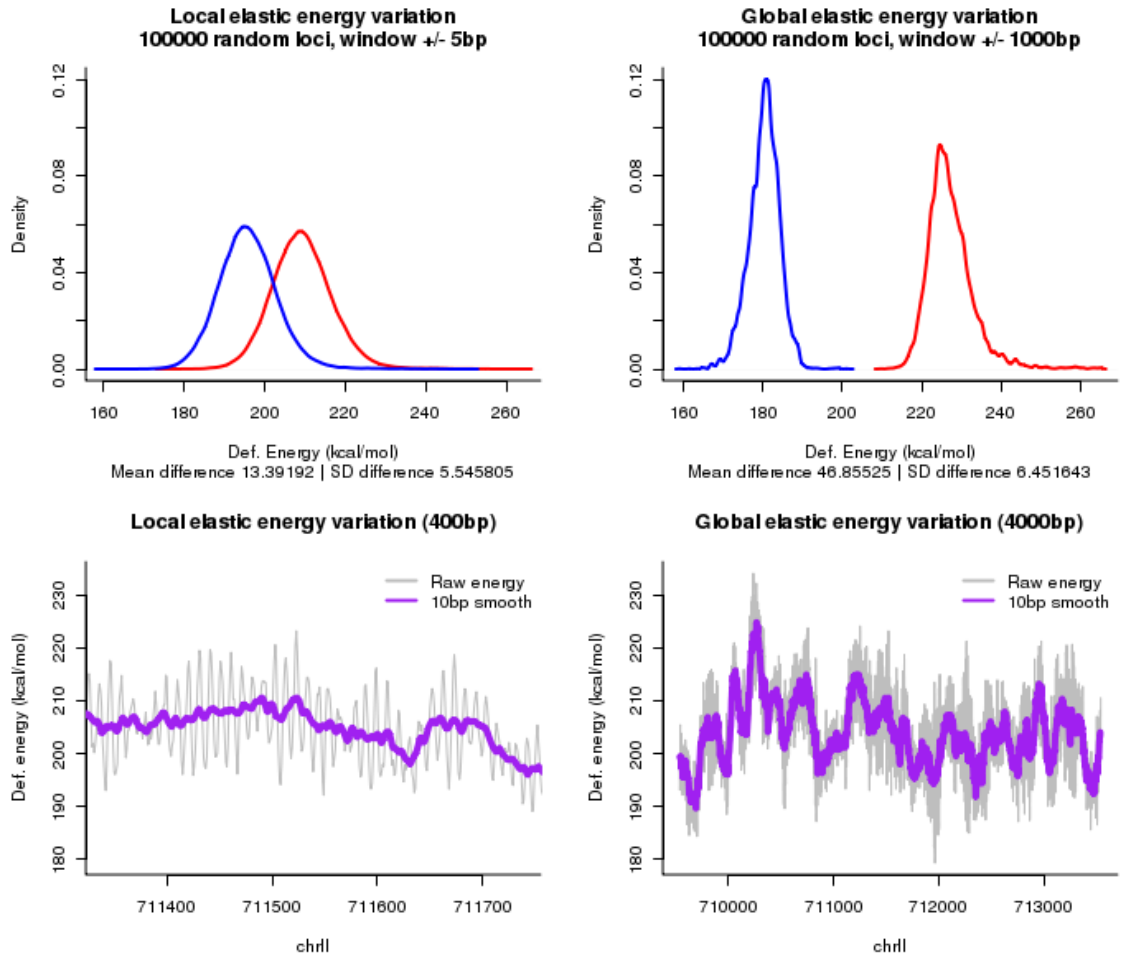
Supplementary Figure S1. Flow cytometry analysis and fluorescence microscope images of late G1 synchronized cells (upper panel) and asynchronous cells (lower panel) for *replicas* 1 (left) and 2 (right).



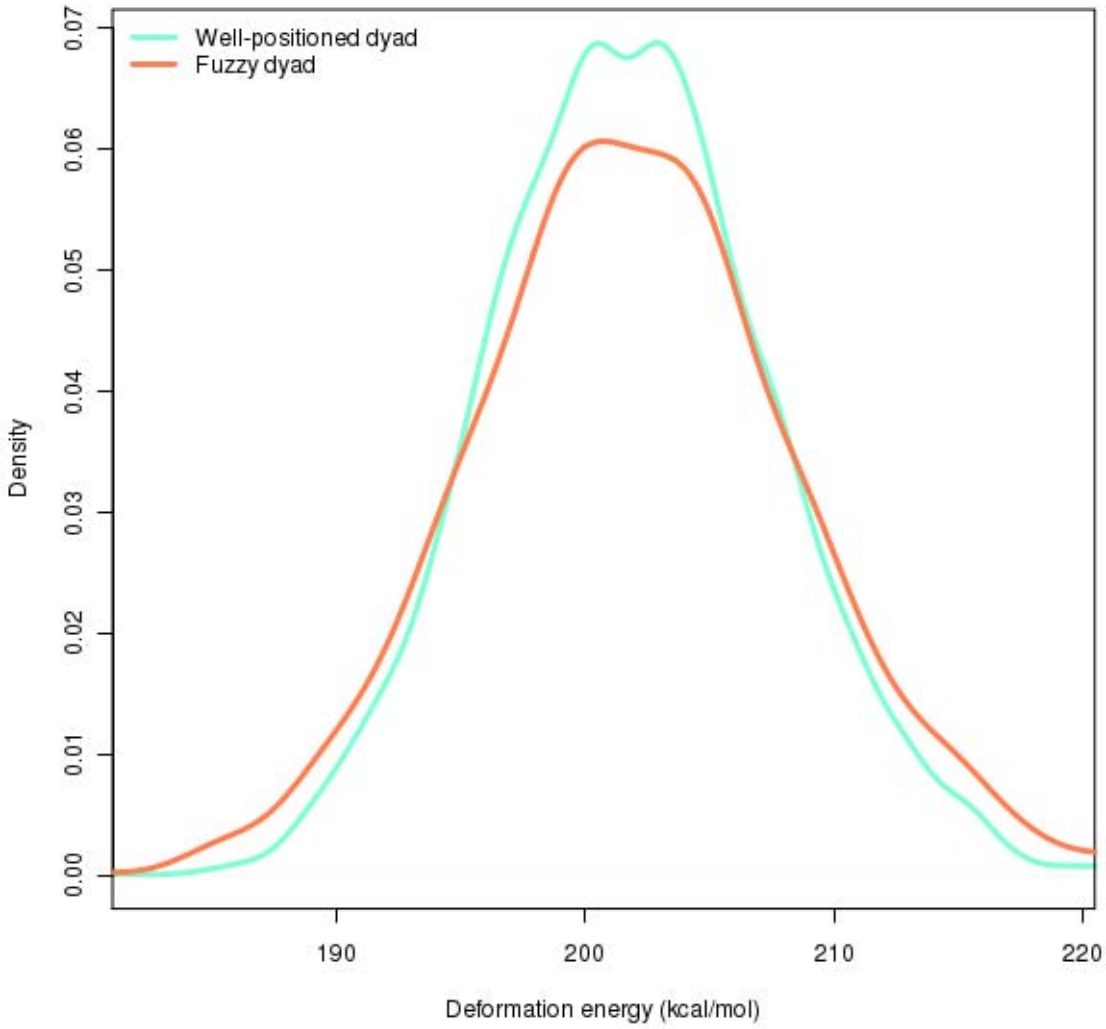
Supplementary Figure S2. MNase digestion profiles of replica 1 (top-left), replica 2 (top-rigth), over-digested sample (bottom-left) and under-digested sample (bottom-right). The left panels show the size distribution of digested DNA molecules as measured by Bioanalyzer and the right panels show the agarose gel analysis of digestion products.



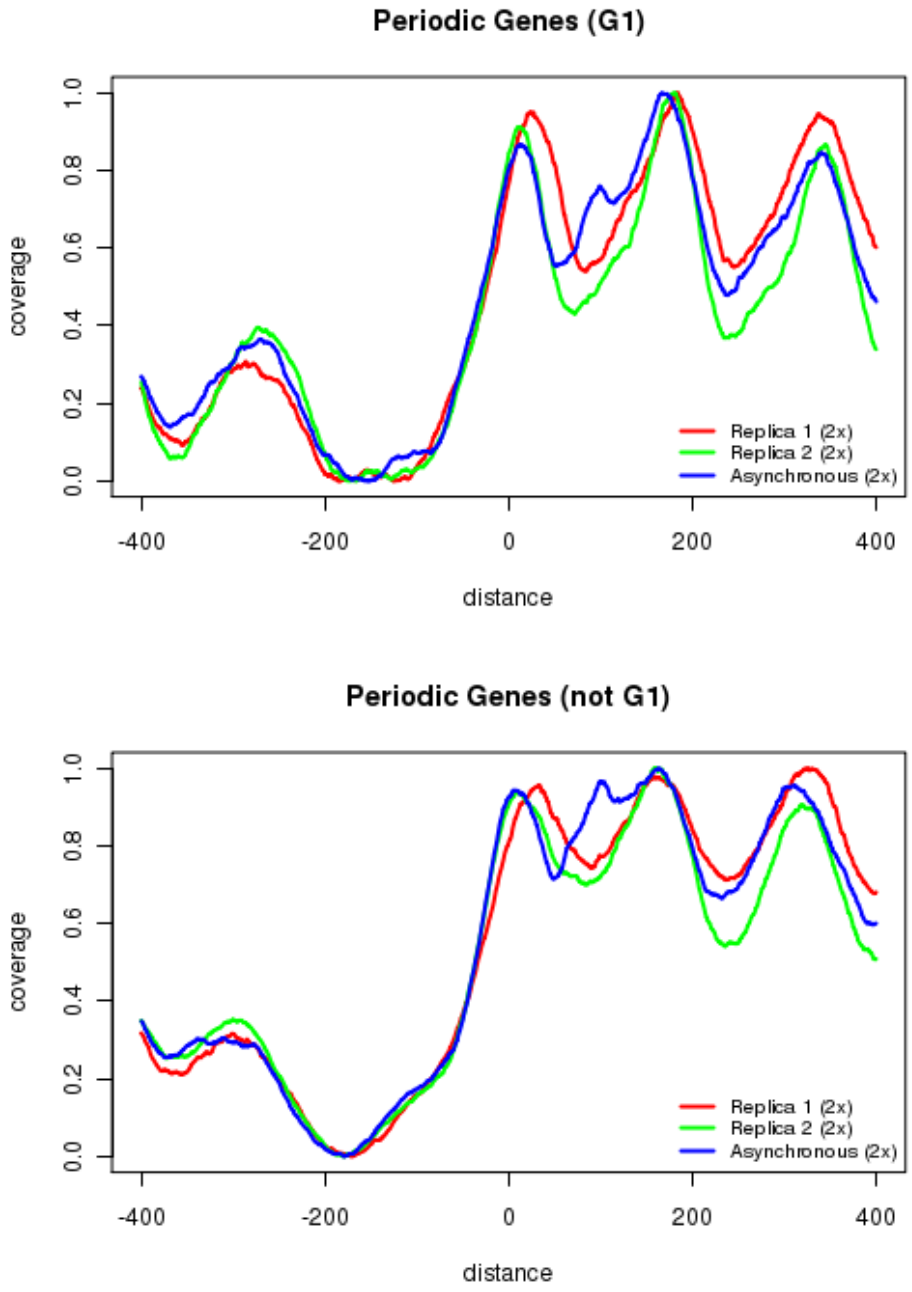
Supplementary Figure S3. Gene clustering according to nucleosomal architecture at transcription start sites. Pie-chart shows the gene distribution for the most populated classes in the sample Replica 2 (2x). For every class, an example of the nucleosome coverage around the TSS of a representative gene is illustrated (window -300:300 from the TSS, marked in red). All plots show the coverage in 5'->3' direction, representing the +1 nucleosome as the peak overlapping or immediately downstream TSS and the -1 nucleosome as the peak right upstream of +1. In the case of SLC1, -1 nucleosome peak is not detected in the -300:300 window.



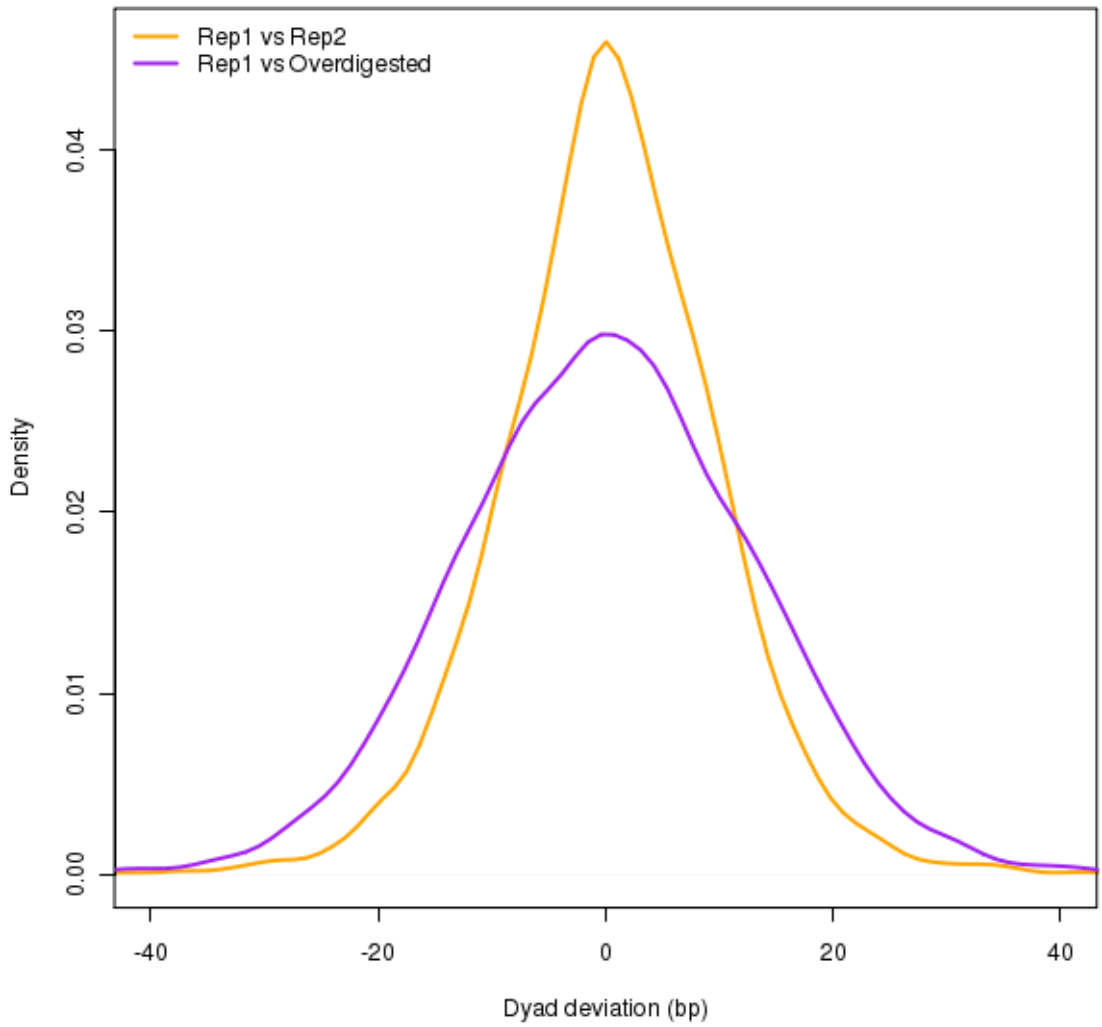
Supplementary Figure S4. Local and global energy variation. Despite local energy variation involves a strong periodicity of 10bp with small energy fluctuations (around 13.39 kcal/mol) (left), these don't act as a strong regulator of the nucleosome fuzziness. Global energy barriers with a larger mesoscopic effect (around 46.86 kcal/mol) could act as intrinsic regulator of the nucleosome phasing along different cells. On top, minimum (blue) and maximum (red) values in a window of +/- 5bp (left) and +/- 1000bp right of 100000 random loci. On the bottom, we show the raw energy (grey) and the 10bp average (purple) of single random region of the chromosome II (left: 400bp window, right:4000bp window).



Supplementary Figure S5. Deformation energy of well-positioned and fuzzy nucleosomes. Deformation energy around +/-5bp around the peak summit has been calculated for annotated -1/+1 nucleosomes. Mean value of every 10 possible combinations was used to account for local periodicity. Fuzzy and well-positioned nucleosomes are taken from the common ones in *replicas* 1 and 2.



Supplementary Figure S6. Effect of cell-cycle periodic genes in nucleosome map. Coverage of cell-cycle periodic genes is shown for G1 related genes (top, 211 genes) and in other stages (bottom, 365 genes). Asynchronous sample (blue) shows a larger perturbation between the +1/-1 nucleosome peaks in both cases.



Supplementary Figure S7. Comparison of dyad deviations due to different digestion. Dyad distances of annotated -1/+1 nucleosomes (coverage peak summits) have been calculated between biological replicates and over-digested sample. Absolute mean deviation between Rep1 and Rep2 is 14.25 bp, compared with 18.75bp (+4.5bp) in the case of Rep1 with over-digested sample.

5. Hog1 bypasses stress-mediated down-regulation of transcription by RNA polymerase II redistribution and chromatin remodeling

Supplementary material for this work includes:

- Supplementary Figures

Additional tables are available in the editor's website http://dx.doi.org/10.1186/gb-2012-13-11-r106

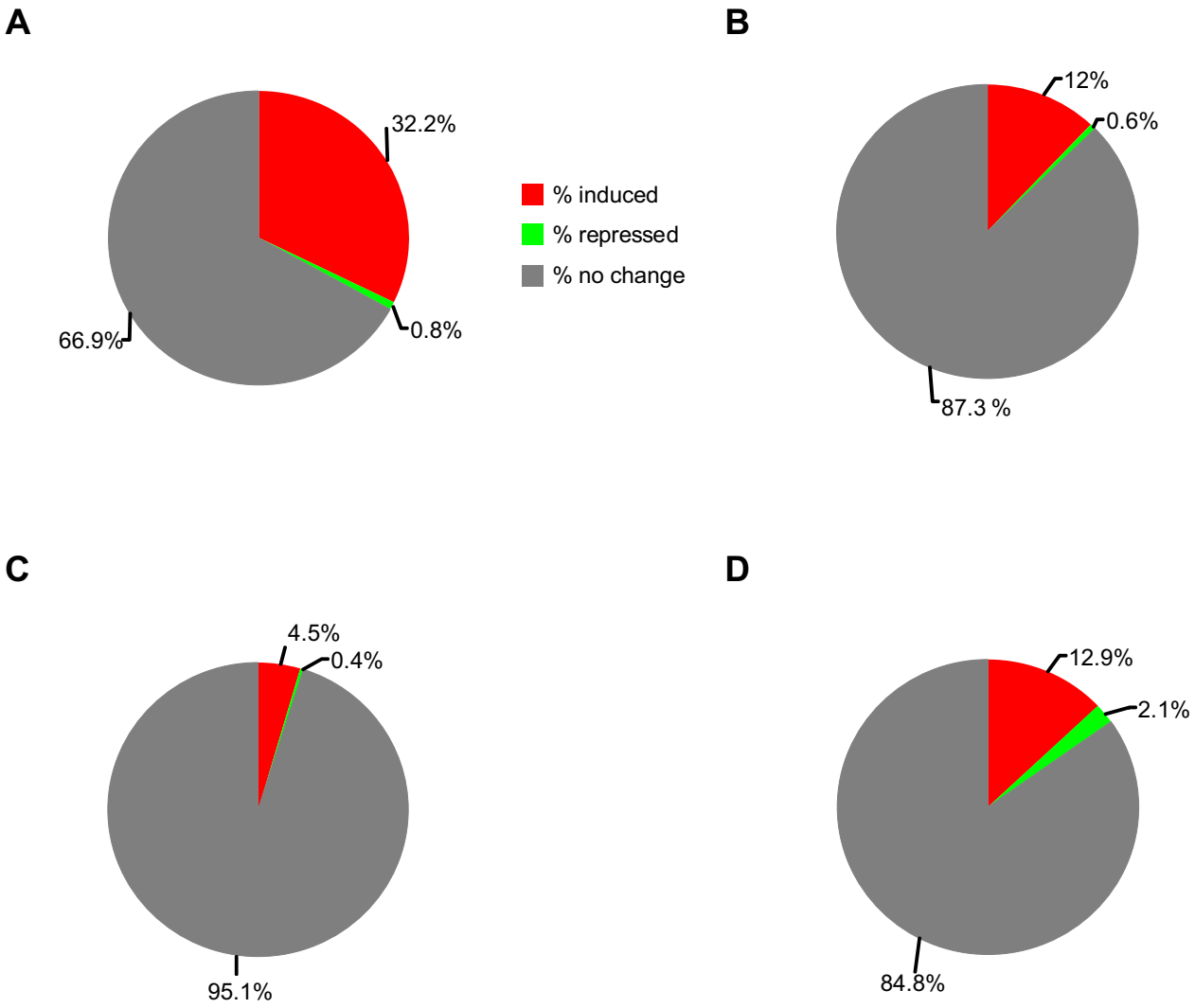


Fig S1 | Expression of osmo-responsive genes remains mostly unchanged upon other environmental stresses. Total osmo-responsive genes (662) were compared to available expression arrays with different environmental stresses (Gasch *et al*., 2000) such as **(A)** heat stress (15 minutes at 37ºC), **(B)** oxidative stress (320 mM H_2O_2 for 30 minutes), **(C)** DTT (250 mM for 60 minutes) and **(D)** aminoacid starvation (30 minutes). Genes were determined to be induced (FC>2), repressed (FC<-2) or unchanged if fold change remained within the induced/repressed threshold upon stress. These genes are shown in red, green and grey respectively together with the percentage from the total genes considered.

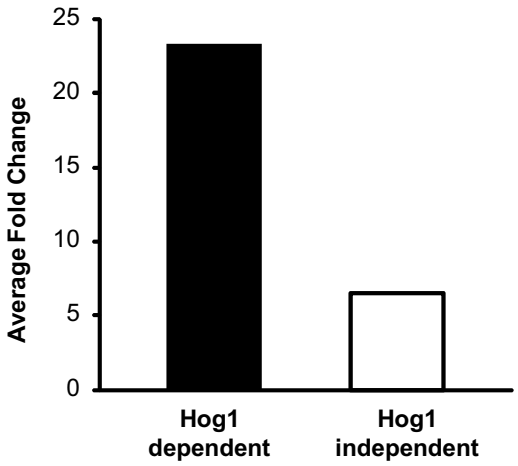


Fig S2 | Hog1-dependent genes have higher expression than Hog1-independent genes upon osmostress. Comparison of average fold change upon stress of Hog1-dependent (black) and independent (white).

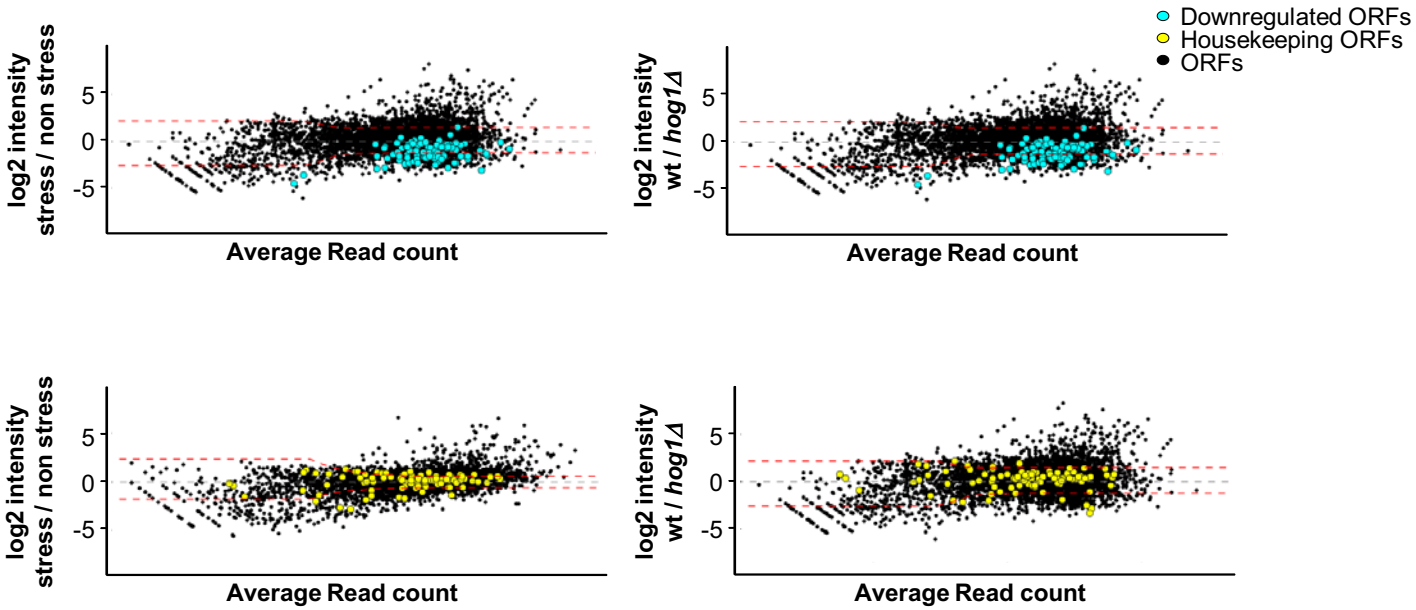


Fig S3 | MA Plots representing RNA Pol II in a subset of 100 genes (blue) whose expression is down-regulated upon more than two fold upon osmostress and a group of 100 genes (yellow) whose fold change upon stress remained unchanged (FC 1 to 1.1) Left panels show wild type RNA Pol II recruitment while right panels represent Hog1-dependent recruitment of RNA Pol II.

A

	Hog1 enrichment	region detail
RNA Pol II	300	
RNA Pol III	18	16 tRNAs 1 SCR1 1 RPR1
LTR	22	

B

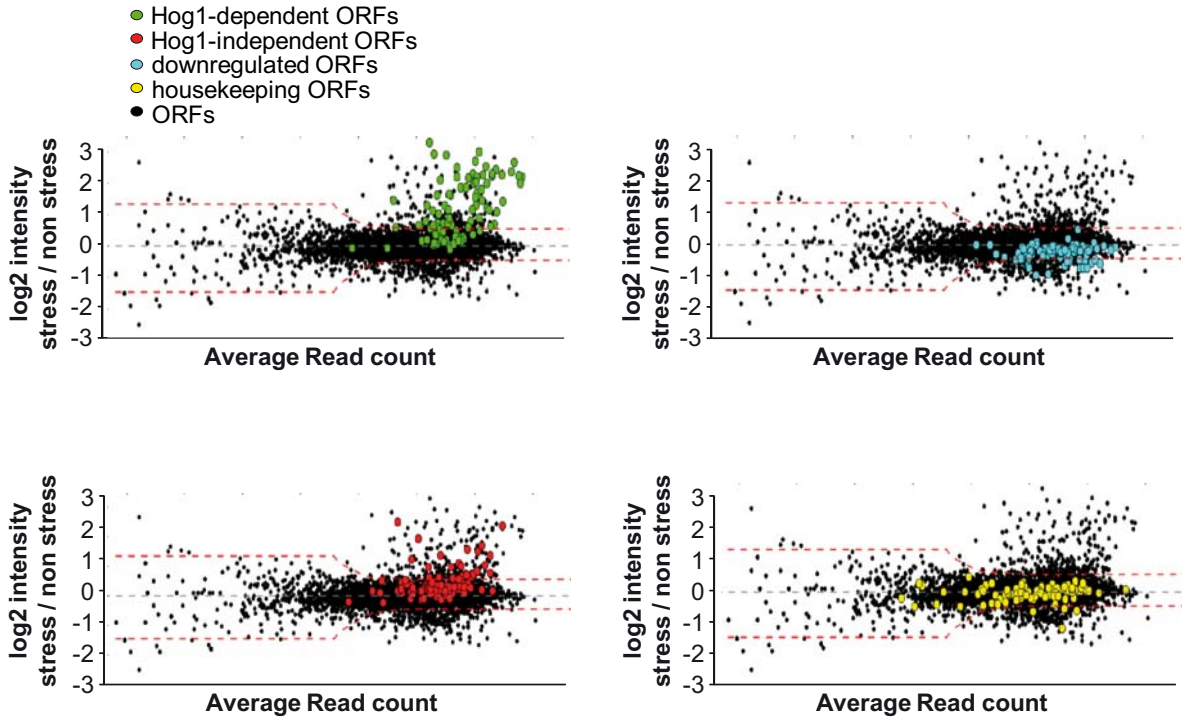
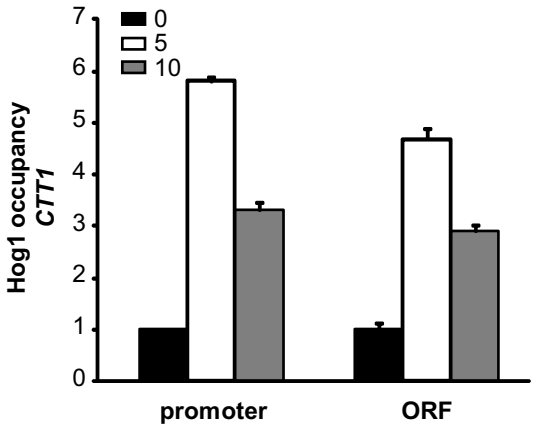


Fig S4 | A. Classification of Hog1 recruited genes or regions depending on their annotation in the SGD as RNA Pol II, RNA Pol III or Long Terminal Repeats (LTR). **B.** MA Plots representing Hog1 binding at the different sets of genes. A group of 100 Hog1-dependent genes are shown in green while 100 Hog1-independent genes are shown in red. In blue and yellow, sets of 100 genes that are downregulated and whose expression remains constant respectively as in Figure S3.

A



B

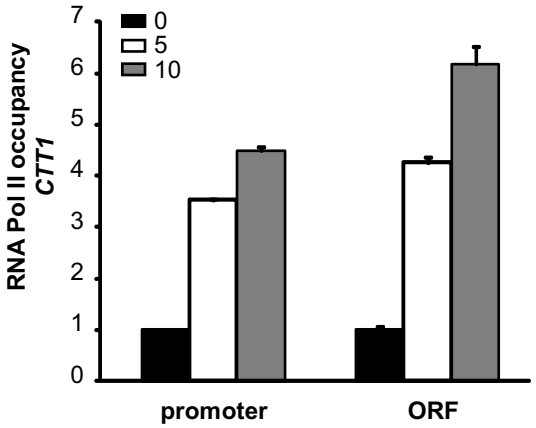


Fig S5 | Binding profile of Hog1 **(A)** and RNA Pol II **(B)** to *CTT1* osmo-responsive gene. Binding of Hog1 and RNA Pol II (Rpb1) validated by chromatin immunoprecipitation as in Figure 3D at the indicated regions of *CTT1* and times upon osmostress (0.4 M NaCl).

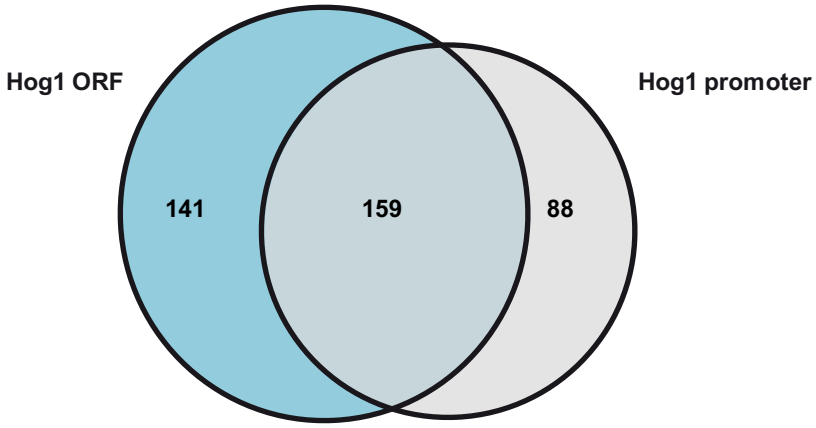


Fig S6 | Genome-wide distribution of Hog1. Venn diagram showing the number of genes in which the ORF and/or promoter (500 bp upstream of ATG) regions were occupied by Hog1 (z-score> 4).

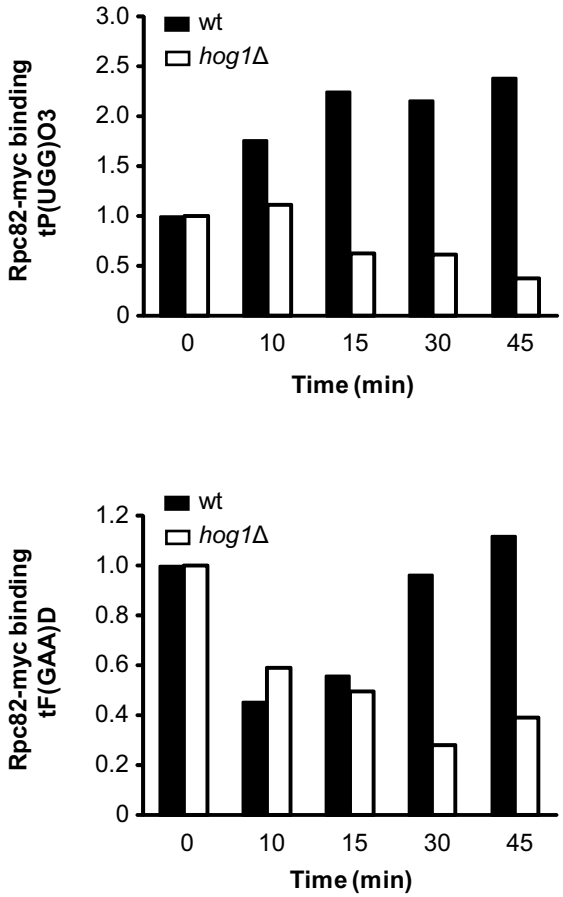


Fig S7 | Association of RNA PolII to tRNAs upon stress depends on *HOG1.* Association of specific subunit of RNA Pol III (Rpc82-myc) was assessed by ChIP upon osmostress (0.4M NaCl) at the indicated times and tRNA loci. Conventional PCR was performed and results are shown as fold induction of treated versus non-treated (time 0) and normalized against internal loading control (*TEL1*).

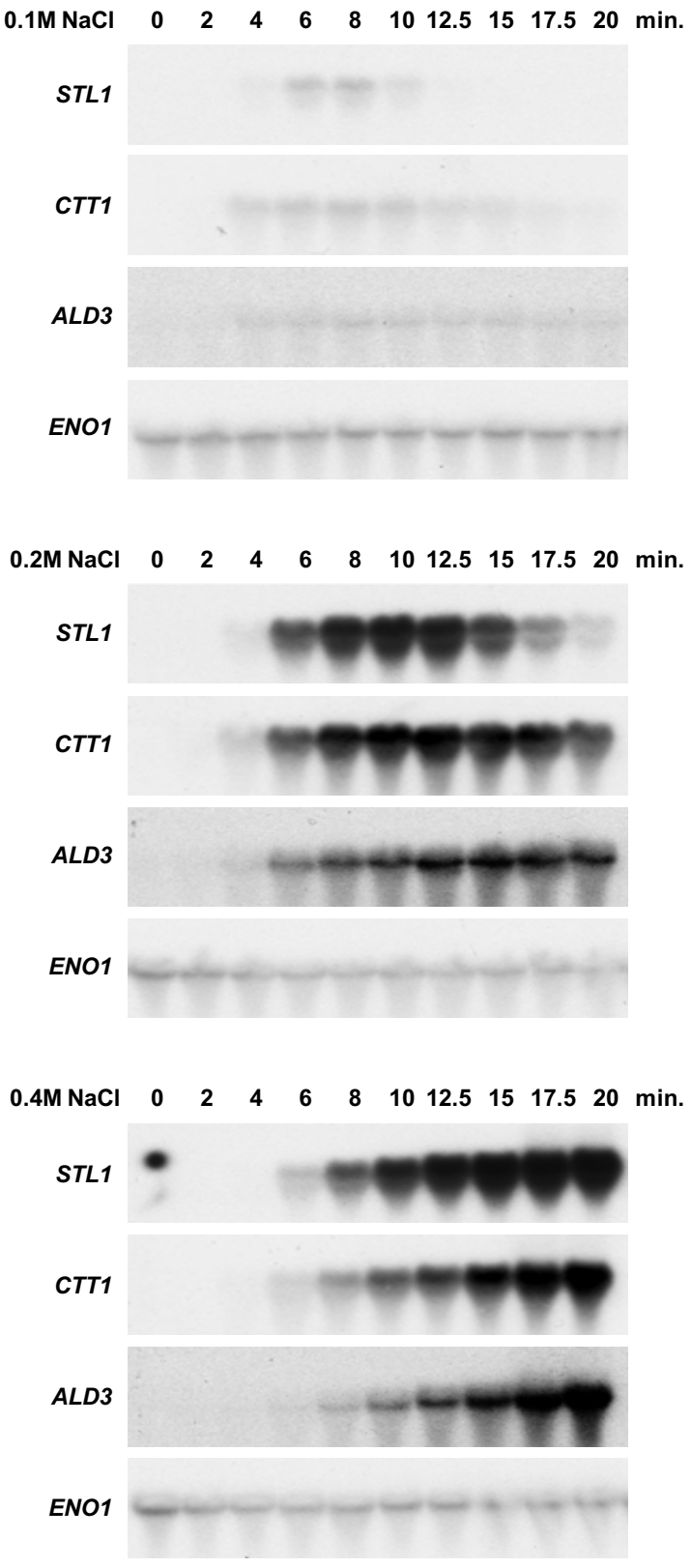


Fig S8 | Kinetics of dose response (0.1M, 0.2M and 0.4M NaCl) expression of different osmostress genes (*STL1*, *CTT1* and *ALD3*). Total RNA was extracted at the indicated times and concentrations. Expression of osmo-responsive genes was probed and normalized with *ENO1* as loading control.

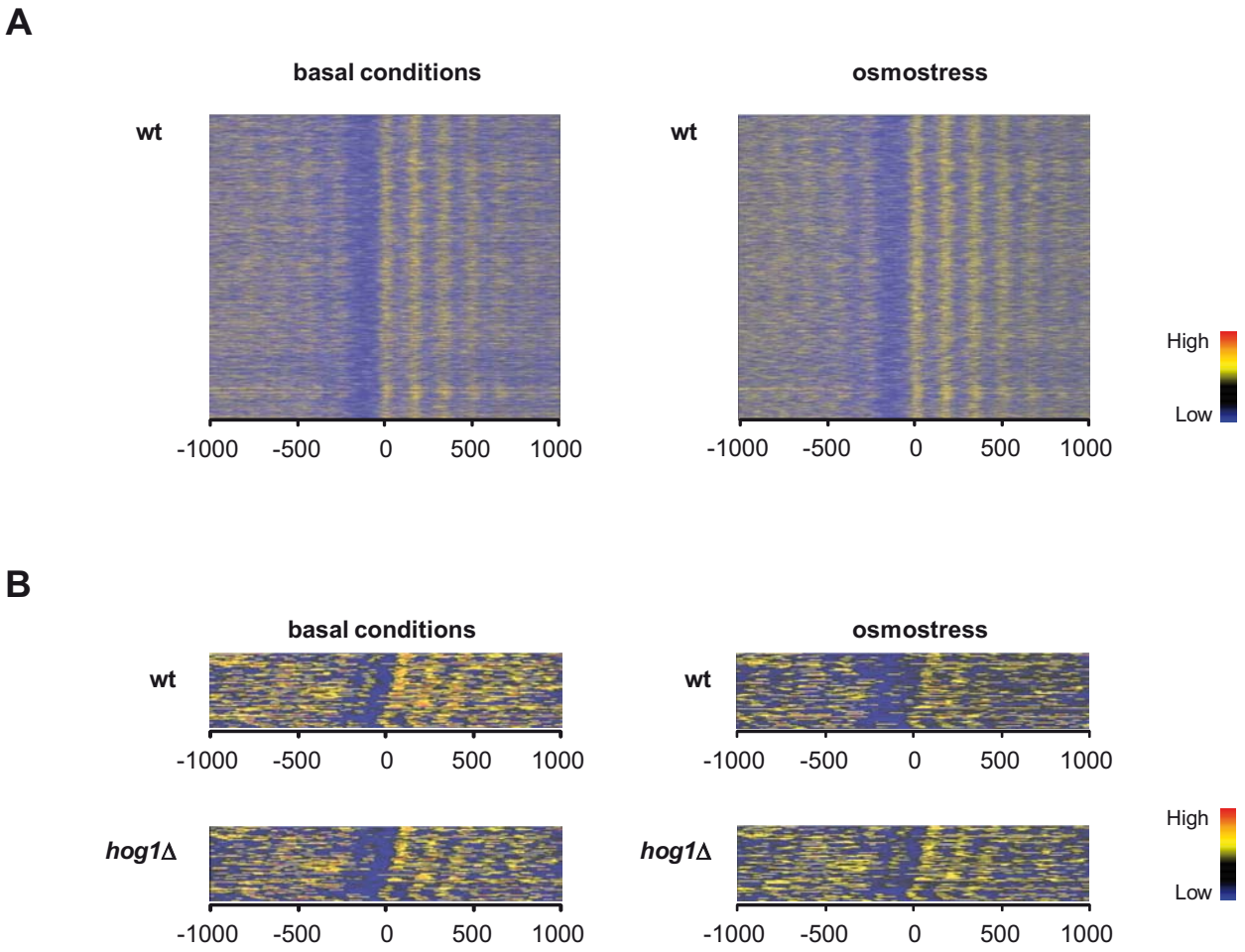


Fig S9 | Nucleosome occupancy maps determined by MNase-Seq genome-wide **(A)** and for Hog1-dependent genes (described in material and methods) in a *wild type* and *hog1* strains **(B)**. Coverage values of each position in the genome has been normalized, converted to reads per million (rpm) and represented in logarithmic scale . Genes (rows) are aligned by their TSS and sorted to maximize the correlation in 0 to 100 bp region. Blue, black, yellow and red indicate a low, medium, and high read density.

6. nucleR: a package for non-parametric nucleosome positioning

Supplementary material for this work includes:

- Supplementary Materials and Methods
- Supplementary Figures
- nucleR's *vignette*

Additional materials are available in the website http://mmb.pcb.ub.es/nucleR/ and in the Bioconductor repository, http://bioconductor.org/packages/release/bioc/html/nucleR.html, including the source code and the Reference Manual

Oscar Flores and Modesto Orozco (2011). nucleR: a package for non-parametric nucleosome positioning. Bioinformatics 2011, doi:10.1093/bioinformatics/btr345

Experimental dataset

For evaluation purposes, we used a nucleosome map obtained from a NGS experiment performed in our group[1]. In the nucleosomal DNA preparation, exponentially growing yeast cells were first cross-linked with formaldehyde, spheroplasted with zymolase and finally subjected to a MNase partial digestion to generate core nucleosomes containing DNA fragments of around 147 bp. Cleaved DNA sample was sequenced on the Illumina/Solexa Genome Analyzer (GA) IIx to generate paired-end reads. Data was processed with standard GA base calling pipeline to convert initial raw images into sequences. The short reads were mapped to *S.cerevisiae* 2003 genome[2]with Bowtie software[3], allowing up to 2 mismatches and discarding those reads mapping in multiple loci. The data was imported into R software and preprocessed using the HtSeqTools[4]package. Resulting reads where processed by nucleR as described in the main text. For further details about the experimental protocol or preprocessment steps see reference [1]. The coverage maps used in the benchmark section is available for download here (RData format, 11Mb)

Strand correction for single ended sequencing reads

In single ended sequencing only one end of the DNA fragment is available, mapping only in one strand. This makes the reads mapped on the positive strand being shifted respect those from the same fragment on the negative strand. In the preprocessment of the NGS reads, nucleR corrects this bias by shifting both strands downstream to overlap reads from both strands, avoiding fake coverage peaks and remarking the nucleosome dyad position. In order to apply this shift, an assumption over the average fragment length is needed. nucleR provides an automatic detection of the fragment length based on the calculation of the optimal shift which gives the best agreement between coverage peaks on both strands. If the user already knows the average length of his/her fragments (from sequencing report, gel electrophoresis, assuming 147bp nucleosome length, etc.) a fixed value can be provided accelerating the preprocessment step. Detailed information about the preprocessment and automatic fragment length detection is available inside man page of R functions processReads() and fragmentLenDetect() respectively.

Fast Fourier Transform implementation and peak calling

nucleR transforms the coverage profile to the Fourier Space and knocks-out the noisy components of the signal before performing the inverse transform. This is done by an efficient method of Fourier Transform called Fast Fourier Transform (FFT)[5]. FFT works best with power of 2 length signals, and can be extremely

inefficient if the profile has a length coinciding with a prime number (not factorizable)[5]. In order to work efficiently with extremely large datasets, our FFT implementation presents some tweaks to dramatically increase the performance, requiring seconds of processing instead of days in worst cases. These tweaks include:

1. Filtering by regions, instead of the whole chromosome at once. This is performed isolating the covered regions by splitting the genome on large uncovered (or unavailable) sites.
2. Splitting the large rows of values into fixed-length, power of 2, vectors, allowing a overlapped window before the chunk *i* and *i+1* to allow convergence of the filtered signal.
3. Padding 0 values at the end of a given chunk to enlarge it till nearest power of 2. Added values have tiny impact on the filtered profile and are removed after transformation.

These tweaks have no impact in the obtained results, being the correlation between the native FFT results and ours very close or equal to 1.

Once the signal is clean, the peak calling can be performed by a simple trend changing detection method over the nucleosome coverage value, i.e., we detect a peak in the position *pi* if value(*pi*) > value(*pi+1).* In tiling arrays, the nucleosome coverage is given by a positive hybridization fluorescence ratio, and in sequencing is calculated by a certain amount of reads piled in a single region. In both cases, a nucleosome will appear as a positive peak in the global landscape.

Due to the larger amount of noise in Tiling Array experiments, we recommend 1% of the components in TA and a 2% in NGS. Thanks to the initial format conversion, this is the only difference in the processing of data from both technologies. The percentage of components selected for knock-out is justified by visual inspection of the power spectrum. See Figure S2 for further information.

The Fast Fourier Transform filter is applied instead of a regular running average because it filters the signal removing the noise instead of just smoothing it. Running averages would need a large window for obtaining smooth data profiles, affecting the shape and the positioning of the peaks in the process. See Figure S3 for an example.

Benchmark

We compared nucleR performance with the two main approaches used in nucleosome calling from high-throughtput experiments: a Hidden Markov Model (HMM) based on Lee et al.[6] and a peak detection based nucleosome caller package available on R/Bioconductor repository, ChIPseqR[7].

nucleR method can be applied in both TA and NGS due the normalization of input data, but the other methods are specific for a single technology, so a three-way comparison with a single dataset is not possible.

To provide objective results about the performance of these alternatives, we used two different approaches: measure the reconstruction of the original map from detected nucleosome calls and the detection of nucleosomes in a synthetic map generated with nucleR

Reconstructing maps from calls

In order to evaluate the accuracy of the nucleosome calls in the 3 compared methods with real data, we performed a full nucleosome detection using genome wide maps and then we measured the correlation between the nucleosome coverage created from the calls and the original map.

We used the dataset and the nucleosome calls provided by Lee et al.[6] when evaluating the performance in Tiling Array experiments where a Hidden Markov Model was applied and a High-Throughtput Sequencing experiment performed in our group when evaluating the peak detection algorithm (ChIPseqR[7]).

In both cases the raw data was preprocessed and imported into R where we applied the steps described in the main text to perform nucleosome detection. In the case of ChIPseqR the calls were calculated applying the *pickPeak()* function in the coverage profile with the same threshold as nucleR (percentile 20 of the coverage).

As the nucleosome calls provided by Lee and colleagues are classified in well-positioned and fuzzy nucleosomes, we translated that to 5 phased nucleosome reads for each well-positioned and a region with a mean coverage of 2 reads for fuzzy nucleosomes. The peaks detected by nucleR and ChIPseqR were converted directly to one nucleosome read with the dyad centered on the detected position. These reconstructed reads were created using IRanges library in R and the reconstructed coverage calculated with the associated *coverage()* function.

The Pearson's correlation between the smoothed Tiling Array hybridization profile and results from the HMM was **0.38**, while the same profile shown a correlation of **0.63** with the reconstructed map from nucleR calls.

For the NGS experiment, the correlation of the reconstructed and the original coverage map was**0.03** for ChIPseqR and **0.29** for nucleR.

Synthetic maps

We implemented in nucleR package the function *syntheticNucMap()* which allows the creation of synthetic coverage or enrichment ratio profiles. The details of the function and the description of the parameters are in the corresponding R manual page. Below there is described a summary of the process:

1. Well positioned nucleosomes are placed periodically leaving some bases of linker DNA between them. The number and the width of nucleosomes as the length of linker DNA segments can be arbitrarily set by the user.
2. A ceirtain number of fuzzy nucleosomes is added randomly on top of the well-positioned ones.
3. For each nucleosome read (well-positioned and fuzzy), a random number of repetitions are created slightly shifted respect the "canonical" position, following the constrains of the user.
4. Additionally, a random map of reads can be created to simulate a naked DNA sample and consequently calculate the ratio between the nucleosome and the random map.

   By creating this synthetic map the user is able to know exactly the positions of the original nucleosomes, accounting with "perfect information" that allows the performance test of the different methods. An example synthetic map is shown in Figure S4

   We created a synthetic map using the following call:

```
syntheticNucMap(wp.num=1000,   fuz.num=200,
wp.del=50,
as.ratio=TRUE, rnd.seed=1)
```

   This call generates a map with 1150 nucleosomes (950 well-positioned and 200 fuzzy) with an associated ratio profile. The internal R random seed is set to "1" to allow reproduction of the results.

   Using this map, we computed the nucleosome calls using a HMM based on Lee et al., ChIPseqR and nucleR. As mentioned previously, HMM expects enrichment ratio values and ChIPseqR expects raw coverage values. To make a fair comparison, we calculated the nucleosome calls using nucleR in both contexts.

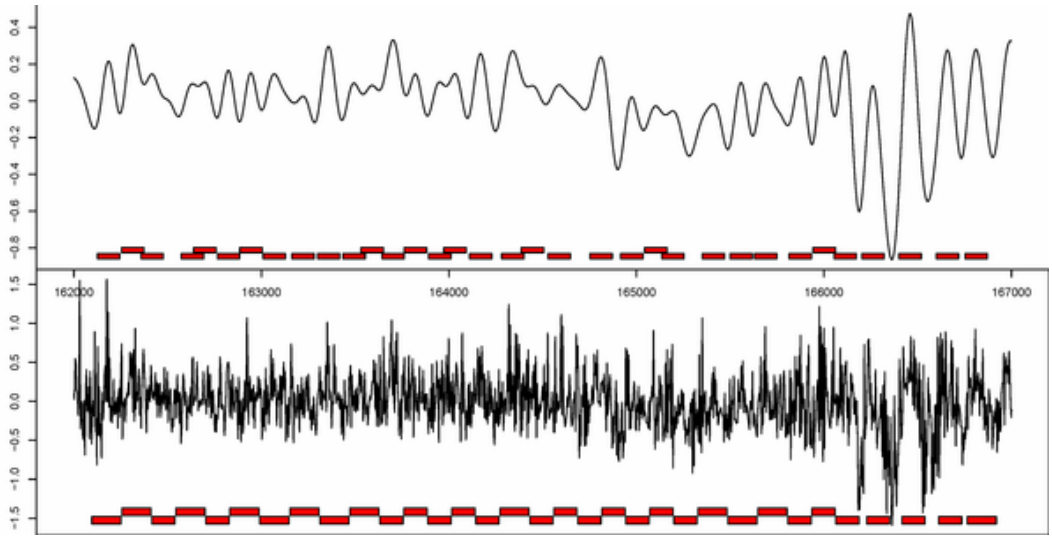
   The following table summarizes the hit percentage (how many nucleosomes where detected) and the mean deviation between the predicted and the real dyad:

	% hits	**deviation (bp)**
nucleR$_{(cover)}$	95.22	2.12
ChIPseqR	81.82	2.24

	% hits	**deviation (bp)**
nucleR$_{(ratio)}$	77.22	13.24
HMM	67.13	26.27

Supplementary Figures

a)



b)

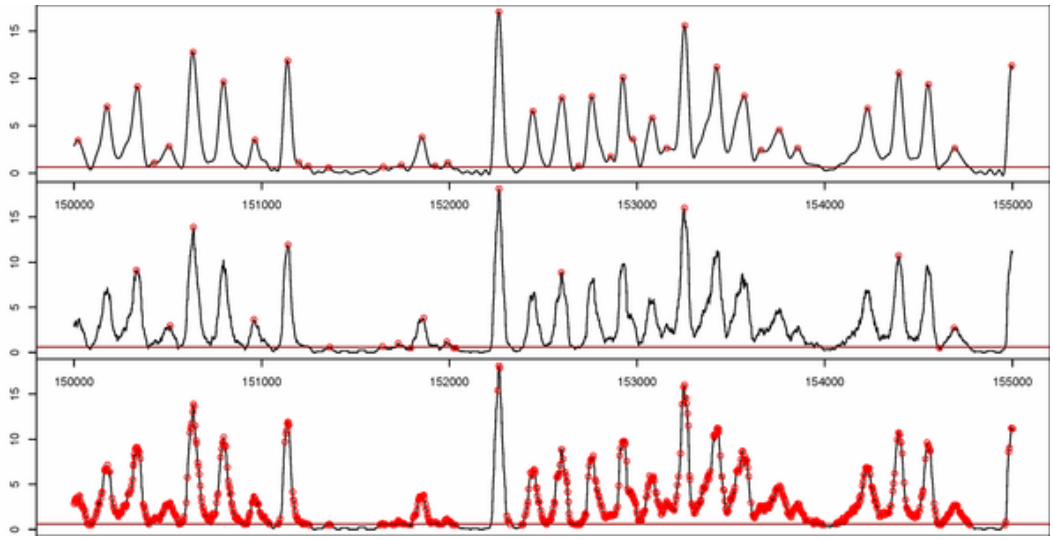


Figure S1. Comparison between presented nucleosome calling methods. **a)** nucleR calls over filtered TA profile (top) versus Lee *et al.* Hidden Markov Model (HMM) calls over raw TA profile (bottom). **b)** nucleR peak detection (red dots) over filtered NGS profile (top) versus ChipSeqR results over raw NGS profile, using default parameters for global maximum (middle) and local maximum (bottom) detection. Used threshold (percentile 25 of the sample) appears as a dark red line in the bottom of each track.

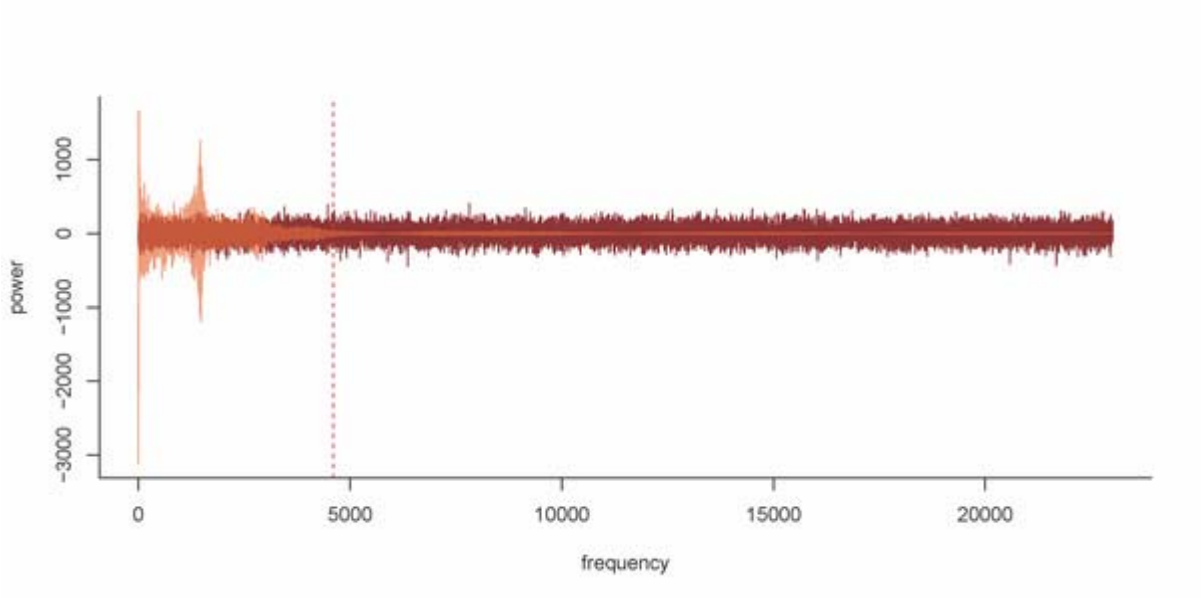
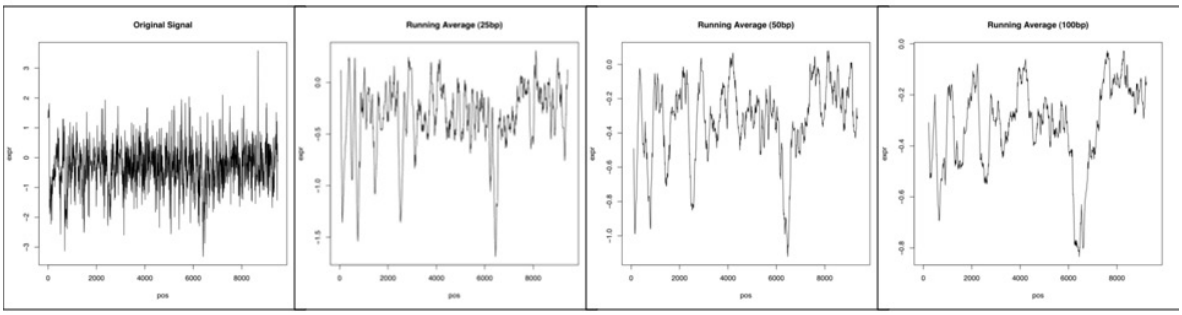


Figure S2. Power spectrum of nucleosomal DNA sample sequenced by NGS (orange) versus a random profile (red). Components at left of dashed red line (percentile 2 of total components) will be the ones saved from the knock-out.

a)



b)

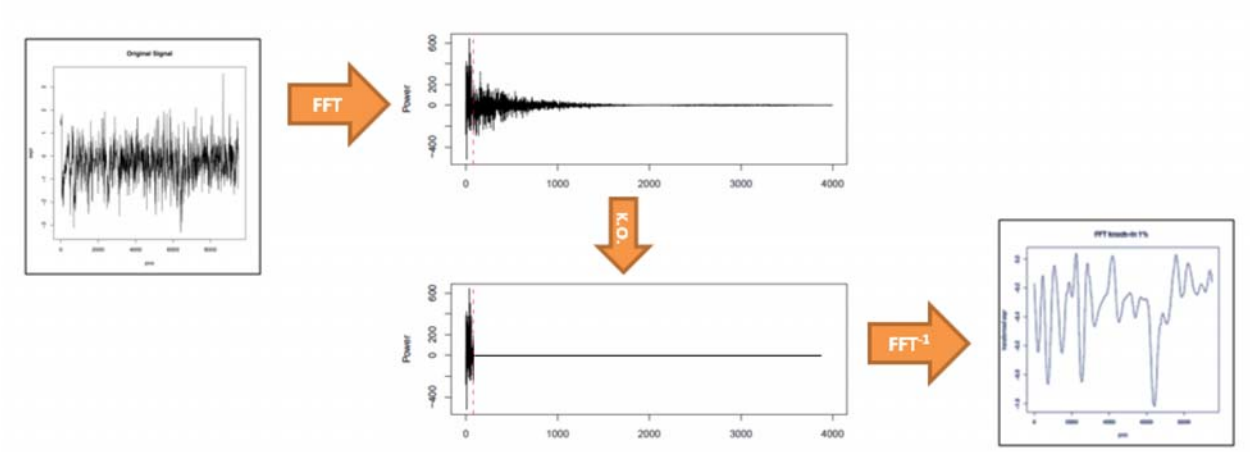


Figure S3. Effect of FFT filtering. **a)** Normal running average smoothing with windows of different length (from left to right: raw data, 25, 50 and 100bp average window). Noise is always present, despite smoothed. **b)** Outline of the filtering process. From raw data the FFT is applied and the components over the marked 1% percentile (red dashed line) are knocked-out setting them to zero. The inverse FFT is applied to obtain the noiseless profile.

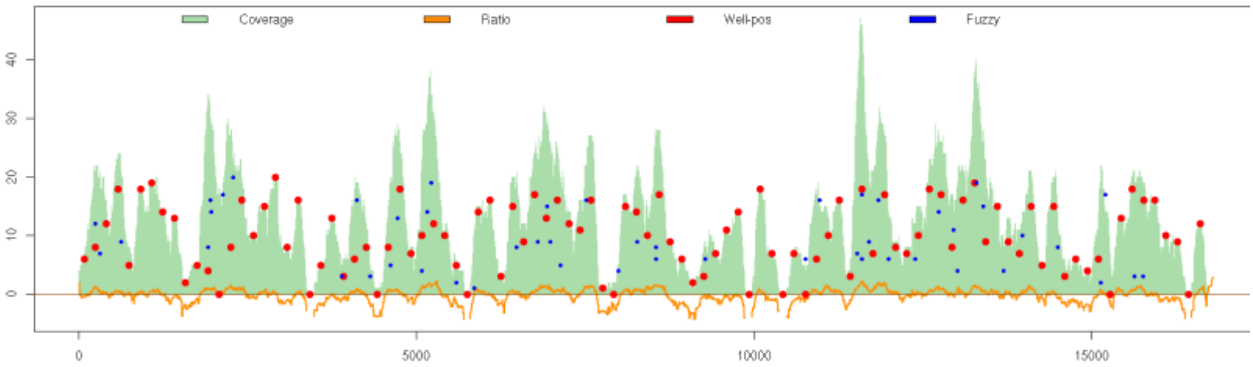


Figure S4. Synthetic nucleosome map generated with nucleR. The coverage profile (green) is calculated from the reads of well positioned (red) and fuzzy nucleosome (blue) (nucleosome dyads represented as dots). The ratio between this coverage and a random sample is represented in bottom (orange).

Supplementary References

1. Deniz, O. et al. (2011) Physical properties of naked DNA influence nucleosome positioning and correlate with transcription start and termination sites in yeast. BMC genomics

2. Downloaded from: ftp://genome-ftp.stanford.edu/pub/yeast/data_download

3. Langmead,B. et al. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology

4. Planet, E. et al. (2011) htSeqTools: High-Throughput Sequencing Quality Control, Processing and Visualization in R. Bioinformatics

5. Smith,S.W. (1999) The Scientist and Engineer's Guide to Digital Signal Processing (Second ed.) California Technical Publishing. Available online: http://www.dspguide.com/6. Lee,W. et al. (2007) A high-resolution atlas of nucleosome occupancy in yeast. Nature genetics7. Humburg, P. (2010) ChIPseqR: Identifying Protein Binding Sites in High-Throughput Sequencing Data. R package version 1.4.0. Available online

Quick analysis of nucleosome positioning experiments using the **nucleR** package

Oscar Flores Guri
Institute for Research in Biomedicine & Barcelona Supercomputing Center
Joint Program on Computational Biology

April 4, 2013

1 Introduction

The nucleR package provides a high-level processing of genomic datasets focused in nucleosome positioning experiments, despite they should be also applicable to chromatin inmunoprecipitation (ChIP) experiments in general.

The aim of this package is not providing an all-in-one data analysis pipeline but complement those existing specialized libraries for low-level data importation and pre-processment into R/Bioconductor framework.

nucleR works with data from the two main high-troughput technologies available nowadays for ChIP: Next Generation Sequencing/NGS (ChIP-seq) and Tiling Microarrays (ChIP-on-Chip).

This is a brief summary of the main functions:

- Data importation: `processReads`, `processTilingArray`
- Data transformation: `coverage.rpm`, `filterFFT`, `controlCorrection`
- Nucleosome calling: `peakDetection`, `peakScoring`
- Visualization: `plotPeaks`
- Data generation: `syntheticNucMap`

For more details about the functions and how to use them refer to the nucleR manual.

This software was published in Bioinformatics Journal. See the paper for additional information.[1]

2 Reading data

As mentioned previously, nucleR uses the pre-processed data of other lower level packages for data importation, supporting a few but common formats that should fulfill the requirements of most users..

`ExpressionSet` from package Biobase [2] is used for Tiling Array experiments as described in Starr [3] and other packages for the Tiling Array manipulation. This kind of experiments can be readed with the `processTilingArray` function.

`AlignedRead` from package ShortRead [4] is recommended for NGS, covering most of the state of the art sequencing technologies. Additionally, support for reads in `RangedData` format is also provided (a range per read with a `"strand"` column) .

2.1 Reading Tiling Arrays

Tiling Arrays are a cheap and fast way to have low-resolution nucleosome coverage maps. They have been widely used in literature[5, 6, 7], but complex statistical methods were needed for their processing [8].

This kind of microarrays cover a part of the genome with certain spacing between probes which causes a drop in the resolution and originates some problems. The nucleosome calling from Tiling Array data required hard work on bioinformatics side and use of heavy and artificious statistical machinery such as Hidden Markov Models [5, 6] or higher order Bayesian Networks [9].

nucleR presents a new method based on a simple but effective peak calling method which achieves a great performance at low computing cost that will be presented in subsequent sections.

In order to standardize the data coming both from Tiling Arrays and NGS, the array fluorescence intensities (usually the ratio of the hybridization of nucleosomal and control sample) are converted to 1bp resolution by inferring the missed values from the neighboring probes. This is done by the function `processTilingArray`:

```
processTilingArray(data, exprName, chrPattern, inferLen=50)
```

An example of a processed dataset is provided in this package. See the help page of `tilingArray_preproc` for details on how it has been created. This object is a numeric vector covering the 8000 first positions of chromosome 1 in yeast (*Saccharomices Cerevisiae* genome (SacCer1)).

```
> require(IRanges)
> library(nucleR)
> data(nucleosome_tiling)
> head(nucleosome_tiling, n=25)
```

```
 [1] 1.273222 1.281978 1.290734 1.299490 1.308246 1.352696 1.397145 1.441595
 [9] 1.486044 1.501795 1.517547 1.533298 1.549049 1.547577 1.546105 1.544633
[17] 1.543161 1.539886 1.536612 1.533337 1.530063 1.488922 1.447782 1.406642
[25] 1.365502
```

This values represent the normalized fluorescence intensity from hybridized sample of nucleosomal DNA versus naked DNA obtained from Starr. The values can be either direct observations (if a probe was starting at that position) or a inferred value from neighboring probes. This data can be passed directly to the filtering functions, as described later in the section 3.

2.2 Next Generation Sequencing

NGS has become one of the most popular technique to map nucleosome in the genome in the last years [10, 11, 12]. The drop of the costs of a genome wide sequencing together with the high resolution coverage maps obtained, made it the election of many scientists.

The package ShortRead allows reading of the data coming from many sources (Bowtie, MAQ, Illumina pipeline...) and has become one of the most popular packages in R/Bioconductor for NGS data manipulation.

A new R package, called htSeqTools [13], has been recently created to perform preprocessing and quality assesment on NGS experiments. nucleR supports most of the output generated by the functions on that package and recommends its use for quality control and correction of common biases that affect NGS.

nucleR handles ShortRead and RangedData data formats. The dataset `nucleosome_htseq` includes some NGS reads obtained from a nucleosome positioning experiment also from yeast genome, following a protocol similar to the one described in [6].

The paired-end reads coming from Illumina Genome Analyzer II sequencer were mapped using Bowtie and imported into R using ShortRead. Paired ends where merged and sorted according the start position. Those in the first 8000bp of chromosome 1 where saved for this example. Further details are in the reference [14]:

```
> data(nucleosome_htseq)
> class(nucleosome_htseq)
```

```
[1] "RangedData"
attr(,"package")
[1] "IRanges"
```

```
> nucleosome_htseq
```

```
RangedData with 18001 rows and 1 value column across 1 space
         space       ranges   |      strand
      <factor>    <IRanges>   | <character>
1         chr1     [1, 284]   |           +
2         chr1     [5, 205]   |           +
3         chr1     [5, 205]   |           +
4         chr1     [5, 209]   |           +
5         chr1     [5, 283]   |           +
6         chr1     [6, 285]   |           +
7         chr1     [6, 285]   |           +
8         chr1     [8, 126]   |           +
9         chr1     [8, 127]   |           +
...        ...          ... ...         ...
17993     chr1 [7994, 8148]   |           +
17994     chr1 [7994, 8150]   |           +
17995     chr1 [7994, 8150]   |           +
```

```
17996     chr1 [7994, 8151]   |          +
17997     chr1 [7994, 8151]   |          +
17998     chr1 [7994, 8151]   |          +
17999     chr1 [7994, 8151]   |          +
18000     chr1 [7994, 8152]   |          +
18001     chr1 [7994, 8152]   |          +
```

Now we will transform the reads to a normalized format. Moreover, as the data is paired-ended and we are only interested in mononucleosomes (which are typically 147bp), we will discard the reads with a length greater than 200bp, allowing margin for some underdigestion but discarding extra long reads. Note that the behaviour of `fragmentLen` is different for single-ended data, see the manual page of this function for detailed information.

As our final objective is identifying the nucleosome positions, and nucleR does it from the dyad, we will increase the sharpness of the dyads by removing some bases from the ends of each read. In the next example, we will create two new objects, one with the original paired-end reads and another one with the reads trimmed to the middle 40bp around the dyad (using the `trim` argument).

```
> #Process the paired end reads, but discard those with length > 200
> reads_pair = processReads(nucleosome_htseq, type="paired", fragmentLen=200)
> #Process the reads, but now trim each read to 40bp around the dyad
> reads_trim = processReads(nucleosome_htseq, type="paired", fragmentLen=200, trim=40)
```

The next step is obtain the coverage (the count of how many reads are in each position). The standard IRanges package function `coverage` will work well here, but it is a common practice to normalize the coverage values according to the total number of short reads obtained in the NGS experiment. The common used unit is *reads per milon* (r.p.m.) which is the coverage value divided by the total number of reads and multiplied per one milion. A quick and efficient way to do this with nucleR is the `coverage.rpm` function. [1]

```
> #Calculate the coverage, directly in reads per million (r.p.m)
> cover_pair = coverage.rpm(reads_pair)
> cover_trim = coverage.rpm(reads_trim)
```

In Figure 1 we can observe the effect of `trim` attribute plotting both coverages. Note that the coverages are normalized in the range 0–1:

2.3 MNase bias correction

The Microccocal Nuclease is a widely used enzyme that has been proved to have a biase for certain dinucleotide steps [14]. In this package we offer a quick way to inspect the effect of such artifact by correcting the profiles of nucleosomal DNA reads with a mock sample of naked DNA digested with MNase.

[1] Note that conversion in the example dataset gives huge values. This is because r.p.m. expects a large number of reads, and this dataset is only a fraction of a whole one. Also take into account that reads from single-ended (or trimmed reads) and reads from paired-ended could have different mean value of coverage

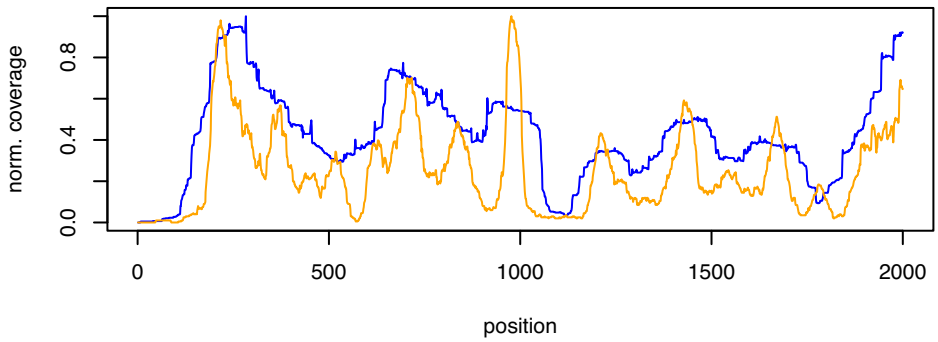


Figure 1: Variation in the sharpness of the peaks using `trim` attribute. In blue, the original coverage; in orange the trimmed version

The use of this function requires a paired-end control sample and a paired end or extended single-read nucleosomal DNA sample. A toy example generated using synthetic data can be found in Figure 2.

```
> #Toy example
> map = syntheticNucMap(as.ratio=TRUE, wp.num=50, fuz.num=25)
> exp = coverage(map$syn.reads)
> ctr = coverage(map$ctr.reads)
> corrected = controlCorrection(exp, ctr)
```

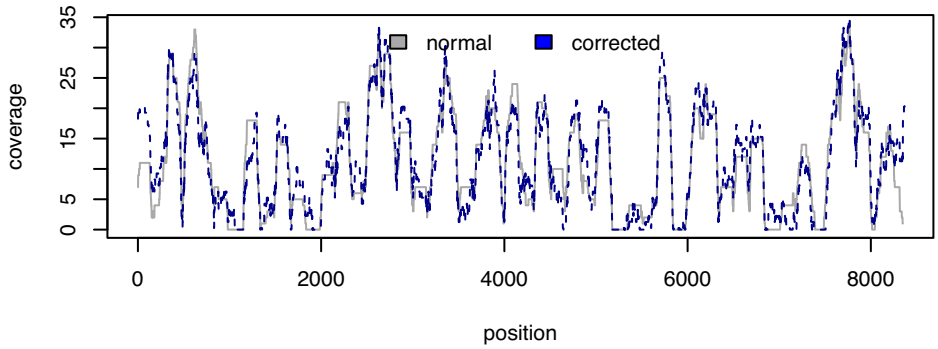


Figure 2: Toy example of MNase biase correction. Random nucleosomal and control reads have been generated using `synteticNucMap` function and corrected using `controlCorrect`

3 Signal Smoothing and Nucleosome Calling

In the previous sections we converted the experimental data from NGS or Tiling Arrays to a continous, 1bp resolution signal. In this section we will remove the noise present in the data and score the peaks identified, giving place to the nucleosome calls.

Previously, in the literature, Hidden Markov Models, Support Vector Machines or other complex intelligent agents where used for this task [5, 6, 9, 15, 12]. This was needed for dealing with the noise and uncertain characterization of the fuzzy positioning of the nucleosomes.

Despite this approach is a valid way to face the problem, the use of such artificious constructs is difficult to implement and sometimes requires a subjective modeling of the solution, constraining or at least conditioning the results observed.

The method presented here proposes to *keep it simple*, allowing the researcher to study the results he or she is interested *a posteriori*.

nucleR aim is to evaluate where the nucleosomes are located and how accurate that position is. We can find a nucleosome read in virtually any place in the genome, but some positions will show a high concentration and will allow us to mark this nucleosome as **well-positioned** whereas other will be less phased giving place to **fuzzy** or **de-localized** nucleosomes [16].

We think it's better to provide a detailed but convenient identification of the relevant nucleosome regions and score them according to its grade of fuzziness. From our point of view, every researcher should make the final decision regarding filtering, merging or classifying the nucleosomes according its necessities, and nucleR is only a tool to help in this *"dirty"* part of the research.

3.1 Noise removal

NGS and specially Tiling Array data show a very noisy profile which complicates the process of the nucleosome detection from peaks in the signal. A common approach used in the literature is smooth the signal with a sliding window average and then use a Hidden Markov Model to calculate the probabilities of having one or another state.

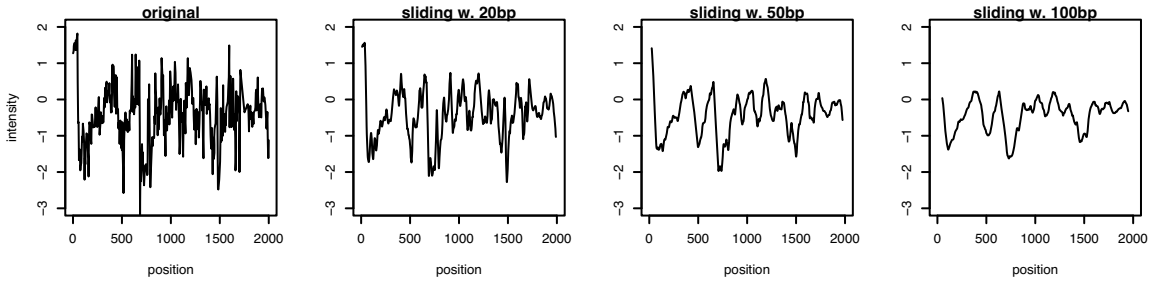


Figure 3: Original intensities from tiling array experiment. Smoothing using a sliding window of variable length (0, 20, 50 and 100 bp) is presented.

As can be seen in Figure 3, data needs some smoothing to be interpretable, but a simple sliding window average is not sufficient. Short windows allow too much noise but larger ones change the position and the shape of the peaks.

nucleR proposes a method of filtering based on the Fourier Analysis of the signal and the selection of its principal components.

Any signal can be described as a function of individual periodic waves with different frequencies and the combination of them creates more complex signals. The noise in a signal can be described as a small, non periodic fluctuations, and can be easily identified and removed [17].

nucleR uses this theory to transform the input data into the Fourier space using the Fast Fourier Transform (FFT). A FFT has a real and a imaginary component. The representation of the real component it's called the power spectrum of the signals and shows which are the frequencies that have more weight (power) in the signal. The low frequency components (so, very periodic) usually have a huge influence in the composite signal, but its rellevance drops as

the frequency increases.

We can look at the power spectrum of the example dataset with the following command:

```
> fft_ta = filterFFT(nucleosome_tiling, pcKeepComp=0.01, showPowerSpec=TRUE)
```

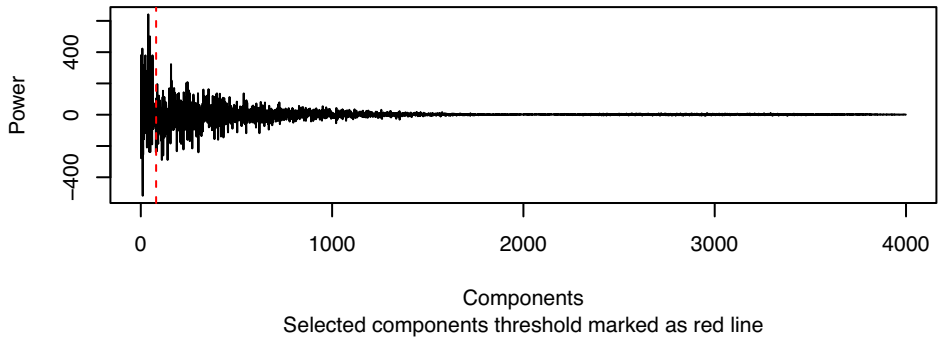


Figure 4: Power spectrum of the example Tiling Array data, percentile 1 marked with a dashed line

In the Figure 4 only the half of the components are plotted, as the spectrum is repeated symmetrically respect to its middle point. The first component (not shown in the plot), has period 1, and, in practice, is a count of the lenght of the signal, so it has a large value.

High frequency signals are usually echoes (repeating waves) of lower frequencies, i.e. a peak at 10 will be the sum of the pure frequence 10 plus the echo of the frequency 5 in its 2nd repetition. Echoes can be ignored without losing relevant information.

The approach nucleR follows is supposing that with just a small percentage of the components of the signal, the input signal can be recreated with a high precision, but without a significant amount of noise. We check empirically that with 1% or 2% of the components (this means account 1 or 2 components for each 100 positions of the genomic data) it's enough to recreate the signal with a very high correlation (>0.99). Tiling Array could require more smoothing (about 1% should be fine) and NGS has less noise and more components can be selected for fitting better the data (about 2%), See Figure 4 for the selected components in the example.

In order to easy the choice of the `pcKeepComp` parameter, nucleR includes a function for automatic detection of a fitted value that provides a correlation between the original and the filtered profiles close to the one specified. See the manual page of `pcKeepCompDetect` for detailed information.

In short, the cleaning process consists on converting the coverage/intensity values to the Fourier space, and knock-out (set to 0) the components greater than the given percentile in order to remove the noise from the profile. Then the inverse Fast Fourier Transform is applyied to recreate the filtered signal. In Figure 5 the filtered signal is overlapped to the raw signal.

The cleaning of the input has almost no effect on the position and shape of the peaks, mantaining a high correlation with the original signal but allowing achieve a great performance with a simple peak detection algorithm:

```
> tiling_raw = nucleosome_tiling
> tiling_fft = filterFFT(tiling_raw, pcKeepComp=0.01)
```

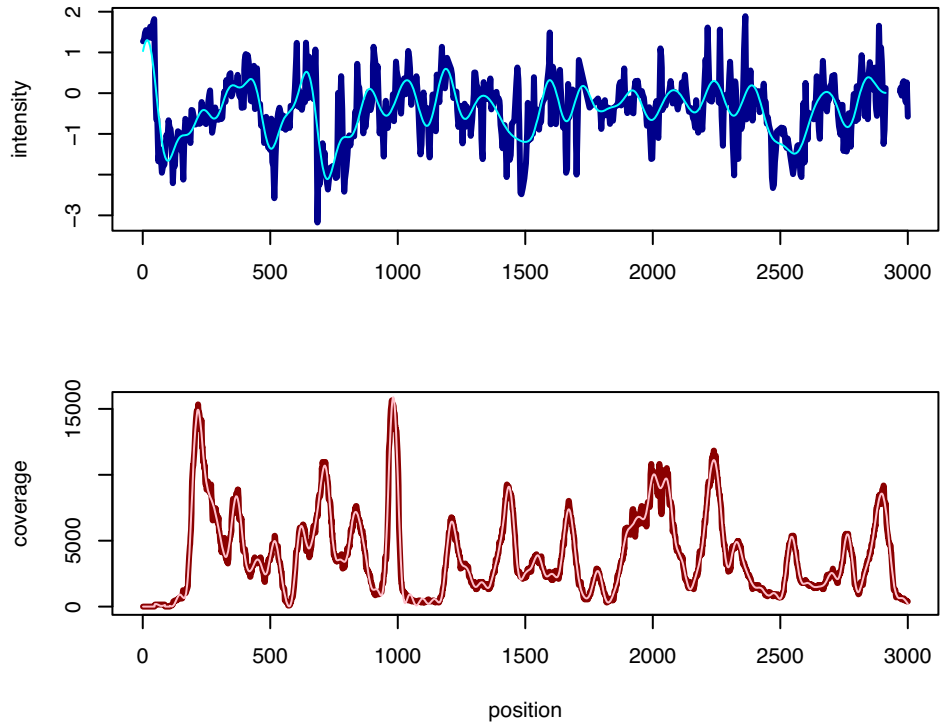


Figure 5: Filtering in Tiling Array (up, blue) (1% comp.) and NGS (down, red) (2%comp.)

```
> htseq_raw = as.vector(cover_trim[[1]])
> htseq_fft = filterFFT(htseq_raw, pcKeepComp=0.02)
> cor(tiling_raw, tiling_fft, use="complete.obs")
```

```
[1] 0.7153782
```

```
> cor(htseq_raw, htseq_fft, use="complete.obs")
```

```
[1] 0.9937643
```

3.2 Peak detection and Nucleosome Calling

After noise removal, the calling for nucleosomes is easy to perform. In nucleosome positioning, in contrast with other similar experiments like ChIP, the problem for the peaks detection algorithms is deal with the presence of an irregular signal which causes lots of local maxima (i.e., peaks due to noise inside a real peak). Here, we avoid this problem applying the FFT filter, allowing the detection of peaks in a simple but efficient way just looking for changes in the trend of the profile. This is implemented in the `peakDetection` function and results can be represented with the function `plotPeaks`:

```
> peaks = peakDetection(htseq_fft, threshold="25%", score=FALSE)
> peaks
```

```
 [1]  218  368  452  518  623  715  776  836  983 1213 1331 1437 1549 1603 1672
[16] 1785 1945 2005 2053 2241 2330 2546 2605 2702 2764 2899 3124 3285 3354 3518
[31] 3710 3869 4009 4137 4233 4305 4383 4524 4596 4832 4912 4981 5171 5241 5291
[46] 5366 5458 5529 5598 5688 5752 5906 5978 6065 6123 6238 6312 6410 6472 6669
[61] 6834 6892 7002 7048 7100 7152 7314 7405 7457 7545 7615 7684 7775 7932 8042
```

```
> plotPeaks(peaks, htseq_fft, threshold="25%")
```

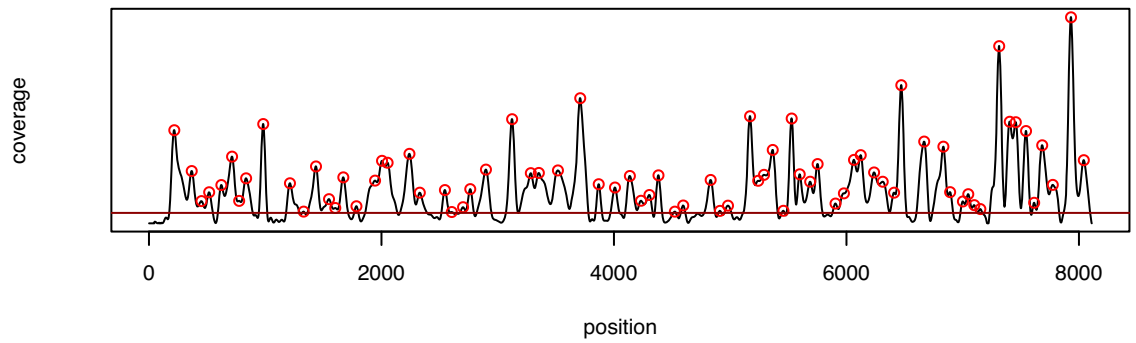


Figure 6: Output of `plotPeaks` function. Peaks are spotted in red and and detection threshold marked with an horitzontal line.

All the peaks above a threshold value are identified. Threshold can be set to 0 for detecting all the peaks, but this is not recommended as usually small fluctuations can apear in bottom part of the profile. This package also provides an automatic scoring of the peaks, which accounts for the two main features we are interested in: the height and the sharpness of the peak.

The *height* of a peak is a direct measure of the reads coverage in the peak position, but represented as a probability inside a Normal distribution.

The *sharpness* is a measure of how fuzzy is a nucleosome. If a peak is very narrow and the surrounding regions are depleted, this is an indicator of a good positioned nucleosome, while wide peaks or peaks very close to each other are probably fuzzy nucleosomes (despite the coverage can be very high in this region).

Scores can be calculated with the `peakScoring` function or directly with the argument `score=TRUE` in `peakDetection`.

```
> peaks = peakDetection(htseq_fft, threshold="25%", score=TRUE)
> head(peaks)
```

```
  peak     score
1  218 0.9749612
2  368 0.6824909
3  452 0.2675856
4  518 0.3829457
5  623 0.4858367
6  715 0.8408881
```

The scores in Figure 7 only account for the punctual height of the peak. As said previously, this measure can be improved by accounting the fuzzyness of a nucleosome call (the sharpness

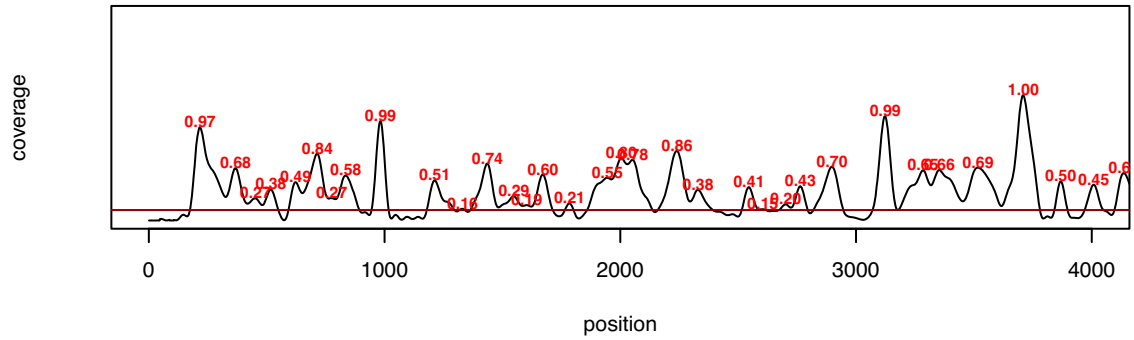


Figure 7: `plotPeaks` function with `score=TRUE`.

of the peak). This requires a way to account for longer range peaks, which can be obtained with the `width` argument. In this way one can convert the identified nucleosome dyads to whole nucleosome length ranges and account for its degree of fuzzyness:

```
> peaks = peakDetection(htseq_fft, threshold="25%", score=TRUE, width=140)
> head(peaks)
```

```
RangedData with 6 rows and 3 value columns across 1 space
     space     ranges |     score   score_w   score_h
  <factor>  <IRanges> | <numeric> <numeric> <numeric>
1        1 [148, 287] | 0.8213296 0.6676980 0.9749612
2        1 [298, 437] | 0.6385168 0.5945427 0.6824909
3        1 [382, 521] | 0.3157506 0.3639156 0.2675856
4        1 [448, 587] | 0.5182733 0.6536008 0.3829457
5        1 [553, 692] | 0.5182437 0.5506507 0.4858367
6        1 [645, 784] | 0.7229812 0.6050742 0.8408881
```

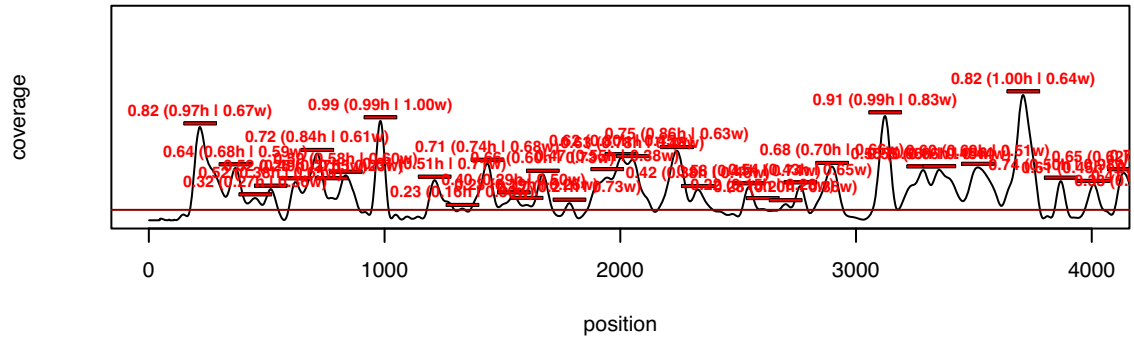


Figure 8: `plotPeaks` output with `score=TRUE` and `width=140`.

Note than in Figure 8 overlapped peaks in a width and tall region are penalized, meanwhile the peaks with surrounding depleted regions have a higher relative score. This is the approach recommended for working with nucleosome calls.

Nucleosome calls filtering, merging or classification can be performed with standard IRanges [18] functions, shuch as `reduce`, `findOverlaps` or `disjoint`.

The next example shows a simple way to merge those nucleosomes which are overlap accounting them as a fuzzy regions:

```
> nuc_calls = ranges(peaks[peaks$score > 0.1,])[[1]]
> red_calls = reduce(nuc_calls)
> red_class = RangedData(red_calls, isFuzzy=width(red_calls) > 140)
> head(red_class)
```

```
RangedData with 6 rows and 1 value column across 1 space
     space       ranges |   isFuzzy
  <factor>    <IRanges> | <logical>
1        1 [ 148,  287] |     FALSE
2        1 [ 298,  905] |      TRUE
3        1 [ 913, 1052] |     FALSE
4        1 [1143, 1854] |      TRUE
5        1 [1875, 2122] |      TRUE
6        1 [2171, 2399] |      TRUE
```

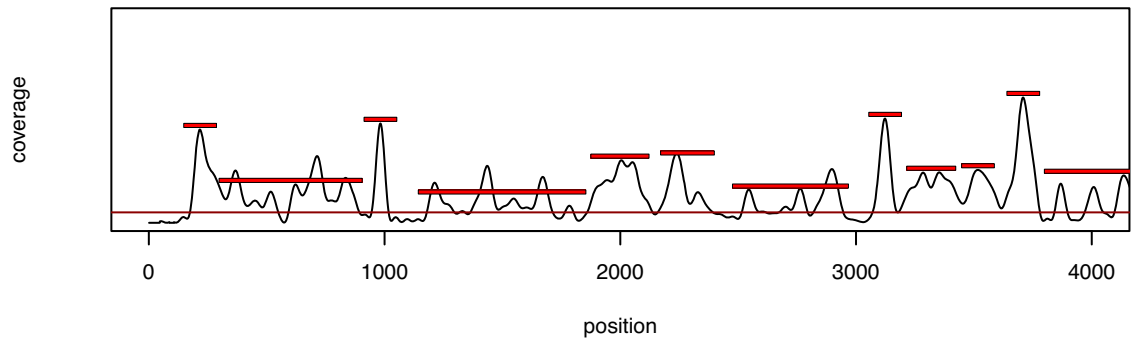


Figure 9: Simple example of ranges manipulation to plot fuzzy nucleosomes

4 Exporting data

`export.wig` and `export.bed` allow exportation of coverage/intensity values and nucleosome calls in a standard format which works on most of the genome browsers available today (like UCSC Genome Browser or Integrated Genome Browser).

`export.wig` creates WIG files wich are suitable for coverage/intensities, meanwhile `export.bed` creates BED files which contain ranges and scores information, suitable for calls.

5 Generating synthetic maps

nucleR includes a synthetic nucleosome map generator, which can be helpful in benchmarking or comparing data against a random map. `syntheticNucMap` function does that, allowing a full customization of the generated maps.

When generating a map, the user can choose the number of the well-positioned and fuzzy nucleosome, as their variance or maximum number of reads. It also provides an option to calculate the ratio between the generated nucleosome map and a mock control of random reads (like a naked DNA randomly fragmented sample) to simulate hybridation data of Tiling Arrays.

The perfect information about the nucleosome dyads is returned by this function, together with the coverage or ratio profiles.

See the man page of this function for detailed information about the different parameters and options.

```
> syntheticNucMap(wp.num=100, wp.del=10, wp.var=30, fuz.num=20, fuz.var=50,
+   max.cover=20, nuc.len=147, lin.len=20, rnd.seed=1, as.ratio=TRUE, show.plot=TRUE)
```

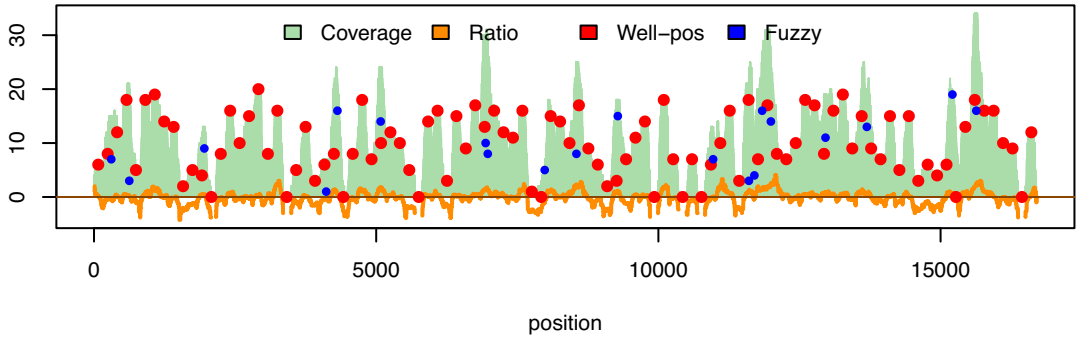


Figure 10: Example synthetic coverage map of 90 well-positioned (100-10) and 20 fuzzy nucleosomes.

References

[1] O. Flores and M. Orozco. nucleR: a package for non-parametric nucleosome positioning. *Bioinformatics*, early publication online:–, 2011.

[2] Robert C Gentleman, Vincent J. Carey, Douglas M. Bates, and others. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004.

[3] Benedikt Zacher, Johannes Soeding, Pei Fen Kuan, Matthias Siebert, and Achim Tresch. *Starr: Simple tiling array analysis of Affymetrix ChIP-chip data*, 2009.

[4] Martin Morgan, Simon Anders, Michael Lawrence, Patrick Aboyoun, HervÃl' PagÃĺs, and Robert Gentleman. ShortRead: a Bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics*, 25:2607–2608, 2009.

[5] Guo-Cheng Yuan, Yuen-Jong Liu, Michael F Dion, Michael D Slack, Lani F Wu, Steven J Altschuler, and Oliver J Rando. Genome-scale identification of nucleosome positions in S. cerevisiae. *Science (New York, N.Y.)*, 309:626–30, 2005.

[6] William Lee, Desiree Tillo, Nicolas Bray, Randall H Morse, Ronald W Davis, Timothy R Hughes, and Corey Nislow. A high-resolution atlas of nucleosome occupancy in yeast. *Nature genetics*, 39:1235–44, 2007.

[7] Travis N Mavrich, Cizhong Jiang, Ilya P Ioshikhes, Xiaoyong Li, Bryan J Venters, Sara J Zanton, Lynn P Tomsho, Ji Qi, Robert L Glaser, Stephan C Schuster, David S Gilmour, Istvan Albert, and B Franklin Pugh. Nucleosome organization in the Drosophila genome. *Nature*, 453:358–62, 2008.

[8] X Shirley Liu. Getting started in tiling microarray analysis. *PLoS computational biology*, 3:1842–4, 2007.

[9] Pei Fen Kuan, Dana Huebert, Audrey Gasch, and Sunduz Keles. A non-homogeneous hidden-state model on first order differences for automatic detection of nucleosome positions. *Statistical applications in genetics and molecular biology*, 8:Article29, 2009.

[10] Noam Kaplan, Irene K Moore, Yvonne Fondufe-Mittendorf, Andrea J Gossett, Desiree Tillo, Yair Field, Emily M LeProust, Timothy R Hughes, Jason D Lieb, Jonathan Widom, and Eran Segal. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, 458:362–6, 2009.

[11] Dustin E Schones, Kairong Cui, Suresh Cuddapah, Tae-Young Roh, Artem Barski, Zhibin Wang, Gang Wei, and Keji Zhao. Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132:887–98, 2008.

[12] Liqun Xi, Yvonne Fondufe-Mittendorf, Lei Xia, Jared Flatow, Jonathan Widom, and Ji-Ping Wang. Predicting nucleosome positioning using a duration Hidden Markov Model. *BMC Bioinformatics*, 11:346, 2010.

[13] David Rossell, Evarist Planet, Camille Stephan-Otto, and Oscar Flores. *htSeqTools: High-Throughput Sequencing Data Analysis*, 2011.

[14] Özgen Deniz, Oscar Flores, Federica Battistini, Alberto Pérez, Montserrat Soler-López, and Modesto Orozco. Physical Properties of Naked DNA Signal Gene Regulatory Regions. *Work under submission*, 2011.

[15] X. Chen, M. M. Hoffman, J. a. Bilmes, J. R. Hesselberth, and W. S. Noble. A dynamic Bayesian network for identifying protein-binding footprints from single molecule-based sequencing data. *Bioinformatics*, 26:i334–i342, 2010.

[16] Cizhong Jiang and B Franklin Pugh. Nucleosome positioning and gene regulation: advances through genomics. *Nature reviews. Genetics*, 10(3):161–72, 2009.

[17] Steven W. Smith. *The Scientist and Engineer's Guide to Digital Signal Processing (Second ed.)*. California Technical Publishing, 1999. Available online: http://www.dspguide.com/pdfbook.htm.

[18] H. Pages, P. Aboyoun, and M. Lawrence. *IRanges: Infrastructure for manipulating intervals on sequences*.

7. htSeqTools: high-throughput sequencing quality control, processing and visualization in R

Supplementary material for this work includes:

- Supplementary Materials and Methods

Additional materials are available in the Bioconductor repository,

http://bioconductor.org/packages/release/bioc/html/htSeqTools.html,

including the source code, the htSeqTools *vignette* and the reference manual

`htSeqTools`: High-Throughput Sequencing Quality Control, Processing and Visualization in R

Evarist Planet [1], Camille Stephan-Otto Attolini [1], Oscar Reina [1], Oscar Flores [2] and David Rossell [1*]

[1]Biostatistics & Bioinformatics Unit, IRB Barcelona, Spain

[2]IRB-BSC Joint Research Program on Computational Biology, IRB Barcelona, Spain

Contents

*to whom correspondence should be addressed

1 Package functionality

`htSeqTools` relies on the general infra-structure set up by Bioconductor (Gentleman *et al*, 2004). Bioconductor provides access to a wide variety of data import, processing and analysis tools, and allows for an efficient implementation of common tasks. Most functions in `htSeqTools` allow for parallel computing by using the `multicore` package (Urbanek, 2011). Table 1 provides running times for several functions when applied to the example in Section 3.1. For instance, computing coverage and 28 pairwise correlations with `cmds` required 274.8 seconds on a single core. The use of 4 cores reduced the time to 114.4 seconds.

We now describe the main functions provided in `htSeqTools`. As much of the emphasis is on quality control and pre-processing, in Section 1.2 we provide a comparison with similar software. We then detail typical workflows for ChIP-seq (Section 1.3) and RNA-seq (Section 1.4) studies. For some detailed examples with R code see Section 3.

1.1 Data import and storage

Several packages are available for importing aligned reads into Bioconductor, *e.g.* `Rsamtools` for BAM (Morgan and Pagés, 2011a), `ShortRead` for BAM, Bowtie, ELAND, MAQ and SOAP (Morgan *et al*, 2011b), and `rtracklayer` for BED, GFF and WIG (Lawrence *et al*, 2011). The input format for `htSeqTools` is based on the data classes defined in the `IRanges` package (Pagés *et al*, 2011). Data from multiple samples are assumed to be stored either in a `RangedDataList` or a `GRangesList` object (`RangedData` or `GRanges` for a single sample). The `RangedData` format is convenient in that a number of Bioconductor packages provide tools to efficiently manipulate and analyze it (*e.g.* coverage computation, overlaps with pre-specified genomic regions etc.). Additionally, the format allows efficient data compression, which greatly reduces the required disk space and time for read/write operations. As an illustration, roughly 15 million reads occupying 800Mb of disk space after Bowtie alignment and bzip2 compression required only 45Mb in the `RangedData` format.

Part of the data reduction in `RangedData` objects is achieved by discarding base calls and quality scores. The information by most `htSeqTools` functions is the chromosome, start and end position that each read was aligned to, and in some cases the strand, which is usually sufficient in studies like ChIP-Seq or RNA-Seq. Notice however that some studies may require additional information, *e.g.* single nucleotide polymorphism (SNP) analysis strategies require base calls and qualities.

	1 core	2 cores	4 cores
`cmds`	274.8 (100%)	191.2 (69.6%)	114.4 (41.6%)
`tabDuplReads`	118.3 (100%)	52.8 (44.6%)	38.4 (32.5%)
`filterDuplReads`	287.5 (100%)	183.9 (64.0%)	161.5 (56.2%)
`islandCounts`	213.7 (100%)	181.6 (85.0%)	157.8 (73.4%)
`enrichedPeaks`	102.2 (100%)	56.1 (54.9%)	36.7 (35.9%)

Table 1: Execution time (seconds) for an increasing number of processors. Obtained on an Apple running OS X 10.6.7 with 2.8GHz processors and 16Gb DDR2 RAM.

1.2 Quality control

We describe the quality control functionality of `htSeqTools`, as well as the following four free softwares: `ShortRead` package in Bioconductor (Morgan *et al*, 2011b); FastQC version 0.9.3; SeqMonk version 0.16.0 (both at `http://www.bioinformatics.bbsrc.ac.uk/projects`); and UCSC Table browser version Oct 15, 2011 (Karolchik *et al* (2004), `http://genome.ucsc.edu/cgi-bin/hgText`). Our aim is not to provide a comprehensive list of all available software and their options, but to illustrate the potential usefulness of `htSeqTools`. In fact, we believe that other software can and should be used in conjunction with `htSeqTools` to better assess data quality.

Table 2 summarizes the software comparison. We consider a block of features related to read quality assessment, which includes inspecting intensities and base-calling accuracy, GC content and detecting adapter sequences. As a second block, we consider the detection of PCR over-amplification artifacts. We evaluate the capacity to detect, statistically assess and remove such artifacts. The third block is based on computing and visualizing correlations between samples, which provides insight into the samples that most closely resemble each other and points to sources of technical bias as well as possible outliers. Finally, we assess the capability to detect problems in sample preparation steps aimed to enrich certain genomic regions (*e.g.* immuno-precipitation) as well as other data pre-processing steps.

After the sequence assessment steps available in `ShortRead`, `htSeqTools` provides several functions that extend the Bioconductor flow. To our knowledge, it is the only software measuring enrichment efficiency and providing a statistical control on the number of read repeats. This procedure allows the user to retain sequences with a few repeats, in situations where a certain number of naturally occurring repeats are expected (*e.g.* short genomes, deeply sequenced samples, experiments targeting a small subset of the genome).

	htSeqTools	ShortRead	FastQC	SeqMonk	UCSC
Sequence assessment					
Base call qualities	No	Yes	Yes	No	No
GC content	No	Yes	Yes	No	No
Adapter detection	No	Yes	Yes	No	No
Over-amplification					
Detection	Yes	Yes	Yes	Yes	No
Statistical assessment	Yes	No	No	No	No
Removal	Yes	Yes	No	R[2]	No
Pre-processing					
Correct strand bias	Yes	No	No	No	No
Extend read length	Yes	No	No	Yes	No
Find read-reach regions	Yes	No	No	Yes	No
Sample correlation					
Compute	Yes	No	No	Yes	R[1]
Visualize	Yes	No	No	Yes	No
Enrichment efficiency	Yes	No	No	No	No

Table 2: Comparison of quality control software functionality. R[1]: restricted to 2 samples and pre-specified genomic regions. R[2]: conserves only reads with no duplicates

1.3 ChIP-seq workflow

Below we indicate the steps in a typical ChIP-seq workflow using `htSeqTools`, as well as the functions needed for each step. Detailed ChIP-seq examples are provided in Sections 3.1-3.2.

1. Detect and remove over-amplification artifacts (`filterDuplReads`, `tabDuplReads`, `fdrEnrichedCounts`)
2. Assess similarity between sample via correlation plots (`cmds`)
3. Extend reads (optional, `extendRanges`)
4. Align peaks (`alignPeaks`)
5. Assess enrichment efficiency (`giniCoverage`, `ssdCoverage`)
6. Find read-rich regions (`enrichedRegions`)
7. Find peaks within regions (`enrichedPeaks`)
8. Annotate peaks (package `ChIPpeakAnno`)
9. Visualize peak location (`PeakLocation`, `stdPeakLocation`, `gridCoverage`)

10. Find regions with accumulation of peaks (`enrichedChrRegions`)

Although here we only include them in Step 2, MDS correlation plots can be used again after read extension, peak alignment or any other desired pre-processing in order to assess their effect on the data. Regarding Step 3, in practice it may be difficult to establish the number of bases that reads should be extended to. In some cases, the extension may actually cause a loss of resolution in peak calling. Hence, we consider Step 3 as optional. We also note that Step 10 is not the primary interest in many ChIP-seq experiments, but included it here for completeness.

1.4 RNA-seq workflow

Here we indicate the basic steps in an RNA-seq workflow. See Section 3.4 for a yeast RNA-seq worked example.

1. Detect and remove over-amplification artifacts (`filterDuplReads`, `tabDuplReads`, `fdrEnrichedCounts`)
2. Assess similarity between sample via correlation plots (`cmds`)
3. Identify expressed RNA *de novo* (`islandCounts`)
4. Compare RNA expression across samples (`enrichedRegions`)
5. Annotate differentially expressed RNA (package `ChIPpeakAnno`)
6. Find regions with accumulation of differential expression (`enrichedChrRegions`)

Similar to the ChIP-seq workflow in Section 1.3, we start by removing over-amplification artifacts. As an important remark, when sequencing short RNAs the same fragment will typically be sequenced many times. That is, one expects a large number of repeats and therefore step 1 (removal of over-amplification artifacts) should be skipped. Next, one inspects if samples cluster appropriately via MDS coverage correlation plots (step 2). Further pre-processing steps are possible. Repeating the MDS correlation plot after pre-processing can help assess whether the processing helped improve the separation between groups.

Steps 3 and 5 are most useful when the genome is poorly annotated, or the user wishes to refine the available annotation, *e.g.* find previously unknown exons. Regarding step 4, many alternatives are available within R for determining differential expression, *e.g.* `edgeR` (Robinson *et al*, 2010),

DESeq (Anders and Huber, 2010), and DEGseq (Wang and Wang, 2011). The approach implemented in enrichedRegions (details in Section 2.3) is specially suited for comparing expression between individual samples. Although it remains useful in datasets with multiple samples per group (see examples in Section 3), in such setups we view it mainly as a quick screening tool and recommend using approaches that formally incorporate the within-group variability (*e.g.* edgeR). Finally, step 6 allows the user to find spatial trends in differential expression, which can suggest common regulatory mechanisms.

2 Detailed method description

2.1 Detecting over-amplification artifacts

High-throughput sequencing requires a PCR amplification step to enrich for adapter-ligated fragments. This step can induce biases as some DNA regions amplify more efficiently than others (e.g. depending on GC content). These PCR artifacts, caused by over-amplification or primer dimers, affect the accuracy of the coverage and can produce biases in downstream analyses. The function filterDuplReads aims to automatically detect and remove these artifacts. The basic rationale is that, by counting the number of times that each read is repeated, we can detect the reads that repeat an unusually large number of times. Ideally, the threshold determining the maximum number of allowable repeats should be determined for each sample separately. The expected number of naturally occurring repeats depends on the genome length, the sequencing depth and the characteristics of each sample. For instance, sequences from IP samples in ChIP-seq experiments focus on a relatively small genomic region while those from controls are distributed along most of the genome, and therefore a higher number of repeats is expected in the former. filterDuplReads determines the threshold in a data-adaptive manner by controlling the False Discovery Rate (FDR). For an example where over-amplified reads were artificially added to a sample see the htSeqTools vignette. For experimental data examples see Section 3.

The basic rationale is that only reads repeating a large number of times are likely to be artifacts. Hence, the null distribution can be estimated by modelling the reads with few repeats. More precisely, let p_j be the proportion of observed reads with j repeats for $j \in \mathbb{N} \setminus \{0\}$, α be a lower bound for the proportion of non over-amplified reads, and q_α be the α quantile associated to p_j. By default we set $\alpha = 0.999$, *i.e.* at most $1/1000$ reads are affected by over-amplification. In statistical terms, our basic assumption is that reads with $1, \ldots, q_\alpha$ repeats are not over-amplification artifacts. We model the

number of repeats for read i, which we denote as X_i for $i = 1, \ldots, n$, as independent realizations from a mixture of Negative Binomial distributions truncated at $1 \leq X_i \leq q_\alpha$. More formally,

$$X_i \sim \sum_{k=1}^{K} \pi_k \text{TNegBin}_{[1,q_\alpha]}(r_k, \theta_k), \tag{1}$$

where K is the number of components in the mixture, and (r_k, θ_k) are the number of failures and success probability in component k. We fit the model via Maximum Likelihood and choose K according to the Bayesian Information Criterion (Schwarz, 1978). Overall, (1) is a flexible model which we observed to fit experimental data reasonably well in a number of scenarios.

In order to estimate the FDR for a given threshold t, we use an empirical Bayes approach similar to that in Efron *et al* (2001). Let $f_0(x)$ be the null distribution for the number of read repeats, $f_1(x)$ the distribution for over-amplified reads and w the proportion of non over-amplified reads. The distribution of X_i can be written as $f(x_i) = wf_0(x_i) + (1-w)f_1(x_i)$, with independence across i. We estimate $f(x)$ with the observed frequencies, imposing that $\hat{f}(x)$ must be monotone decreasing after its mode (using `isoreg` from package `base`) to prevent random fluctuations in $\hat{f}(x)$ for large x. $f_0(x)$ is estimated with the truncated Negative Binomial mixture (1). Setting w to its upper bound 1, an estimated upper bound for the FDR(t) associated to the threshold t is $\frac{\sum_{i \geq t} \hat{f}_0(i)}{\sum_{i \geq t} \hat{f}(i)}$. We enforce that the estimated FDR is monotone decreasing with t via the monotone regression in `isoreg`. By default we set FDR(t) < 0.01.

2.2 Derivation of SD_n and G_n

Denote the number of observed sequences by n. Intuitively, a lack of uniformity in the distribution of such sequences along the genome indicates that some regions were selectively sequenced, *i.e.* enriched in the sample preparation. We measure lack of uniformity by assessing variability in the coverage, i.e. the number of sequences covering each base in the genome. In particular, `ssdCoverage` computes the coverage standard deviation (SD) and `giniCoverage` the Gini index (Gini, 1912), which is a classical econometrics measure of wealth inequality. Because coverage is never uniform in practice (*e.g.* due to sequencing preferences or sequence-dependent biases in sample preparation), strong departures from uniformity are expected in all samples. For this reason, the main usefulness of these indexes lies in comparing variability between a sample of interest and its control.

We first show that, assuming uniformity in the read distribution, the expected value of the coverage SD is proportional to $\sqrt{n}$. Hence, it is preferable to use $\mathrm{SD}_n = \mathrm{SD}/\sqrt{n}$ as a measure that can be compared across samples with different sequencing depths. Let m be the genome length, r the read length (which we assume constant for all reads). and y_k the coverage at position $k \in \{1, \ldots, m\}$. Since each read covers r bases, the probability that an individual read overlaps position k is $\frac{r}{m}$ and the mean coverage is equal to $\bar{y} = \frac{1}{m}\sum_{k=1}^{m} y_k = \frac{nr}{m}$. Assuming independence between reads, $E(y_k) = \frac{nr}{m}$ and $V(y_k) = n\frac{r}{m}(1-\frac{r}{m})$ for all k. The expected value of the coverage variance can thus be computed as

$$E(\mathrm{SD}^2) = E\left(\frac{1}{m}\sum_{k=1}^{m}(y_k - \bar{y})^2\right) = \frac{1}{m}\sum_{k=1}^{m} E\left[\left(y_k - \frac{nr}{m}\right)^2\right] = \frac{1}{m}\sum_{k=1}^{m} V(y_k) = \frac{m}{m}V(y_1) = n\frac{r}{m}\left(1 - \frac{r}{m}\right). \quad (2)$$

Regarding the Gini coefficient G, to our knowledge there is no closed-form expression for its expected value $E(G|n)$ given a number of reads n. Therefore, we resort to a simulation scheme. We repeatedly generate n reads uniformly distributed along the genome and compute the average of the observed G in each simulated dataset. Typically, as the number of reads n is large (tens or hundreds of millions), 1 or 2 simulations are enough to ensure highly precise $E(G|n)$ estimates, which renders the approach computationally feasible. In order to define the n-adjusted Gini (G_n), we notice that the Gini coefficient can be interpreted as the difference between two integrals which compare the observed vs. a uniform cumulative distribution function (cdf). Therefore, we define $G_n = G - E(G|n)$, so that G_n is the difference between the observed and the expected cdf under a sample of n uniformly distributed reads. We note that, although the expected value of SD_n and G_n does not depend on n, their variability will typically increase when n is small. Hence, for small n these indexes are less reliable. A rigorous variability assessment can be performed via Bootstrap.

SD_n is preferable to G_n in terms of computational speed, as it does not require any simulations. The examples in Section 3 explore the performance of both indexes in transcription factor ChIP-seq and chromatin mark ChIP-seq studies. Transcription factor ChIP-seq experiments commonly show a relatively reduced number of strong peaks. While some chromatin marks behave in the same manner, some marks are widely spread and show smaller peaks. The results in Section 3 suggest that, while both indexes perform well for transcription factor ChIP-seq, SD_n is preferable for chromatin mark ChIP-seq. Hence, we recommend SD_n as a default choice.

2.3 Determining rich-read regions and differential expression

In many sequencing experiments it is of interest to screen the genome for regions accumulating a large number of reads. In ChIP-seq studies the screen helps identify and focus on regions with peaks. In RNA-seq studies it may point to previously unknown transcripts, exons or short RNAs. The function `enrichedRegions` implements a computationally efficient scheme to achieve this goal.

We consider setups where the interest lies in a single sample, as well as setups with two or more samples. The basic rationale is to compute the overall read coverage (*i.e.* across all samples under consideration) and select regions with coverage above a user-specified threshold. We refer to such regions as *islands* and denote them as $i = 1, \ldots, I$. Further, let J be the number of samples, x_{ij} the number of reads in sample j overlapping island i, $n_j = \sum_{i=1}^{I} x_{ij}$ the number of reads overlapping some island and $p_{ij} = E(\frac{x_{ij}}{n_j})$. Conditioning on n_j and assuming that reads are independent, the marginal distribution of x_{ij} is $\text{Bin}(n_j, p_{ij})$. This observation suggests a simple approach to test for a significant accumulation of reads in any given island.

In a single sample setup (i.e. $J = 1$), we aim to detect islands with a proportion of reads above average, which can be formalized as testing the null hypothesis $H_0 : p_{i1} \leq \frac{1}{I}$ versus the alternative $H_1 : p_{i1} > \frac{1}{I}$. A straight-forward binomial test is applied to obtain an exact one-sided P-value (`enrichedRegions` also allows the user to choose a two-sided version of the test).

In a setup with ≥ 2 samples, the null hypothesis can be formulated as $H_0 : p_{i1} = \ldots = p_{iJ}$. A P-value is obtained via a likelihood ratio test comparing the null hypothesis with the unrestricted model $p_{i1} \neq \ldots \neq p_{iJ}$. The likelihood-ratio test may be inexact when the expected number of reads in a sample $\frac{\sum_{j=1}^{J} x_{ij}}{J}$ under the null hypothesis is low (as a rule of thumb, below 5). Although such cases are rare, because the islands were defined as regions with high coverage to start with, when they do occur `enrichedRegions` provides the option of using permutation-based chi-square tests.

Finally, the user can select several P-value adjustment procedures to control for multiple testing (*e.g.* Benjamini and Yekutieli (2001)).

2.4 Detecting the accumulation of significant hits

The function `enrichedChrRegions` looks for chromosome regions accumulating a large number of hits, *e.g.* peaks in ChIP-seq or differential expression

in RNA-seq experiments.

Let n be the number of hits and l_i for $i = 1, \ldots, n$ be the location of hit i, given by its chromosome and midpoint between hit start and end. We compute a smoothed number of hits $s(l)$ by counting the number of hits in a moving window of user-specified size (10kb by default). We then report regions with $s(l) > t$, where t is a threshold to control the FDR below a user-specified level (0.05 by default). Similar to the procedure in Section 2.1, an upper bound for FDR(t) is estimated as $\frac{\hat{P}(s(l) \geq t|H_0)}{\hat{P}(s(l) \geq t)}$, where $\hat{P}(s(l) \geq t)$ is the observed number of regions with $s(l) \geq t$. For $\hat{P}(s(l) \geq t|H_0)$ we simulate n hits $l_1^*, \ldots, l_n^*$ uniformly distributed along the genome and compute $s(l^*)$. $\hat{P}(s(l) \geq t|H_0)$ is the mean number of regions with $s(l^*) \geq t$ averaged across several independent simulations.

3 Examples

3.1 Example 1: GSE25836

The aligned reads of the human ChIP-sequencing experiment GSE25836 were downloaded from `www.ncbi.nlm.nih.gov/geo`. For each sample we read the chromosome, strand and start/end positions of the alignments into R via `read.table` and structured them as a `RangedData` objects. Information from all samples was then combined in a `RangedDataList` object named `gse25836` and saved to an .RData file. We start an R session, load the `htSeqTools` and `multicore` packages and the data.

```
> library(htSeqTools)
> library(multicore)
> load('gse25836.RData')
> gse25836
RangedDataList of length 7
names(7): GSM634613 GSM634614 GSM634615 GSM634616 GSM634617 GSM634618
GSM634619
> head(gse25836[[1]])
RangedData with 6 rows and 1 value column across 25 spaces
        space                 ranges |   strand
  <character>              <IRanges> | <factor>
1        chr1 [222562220, 222562251] |        -
2        chr1 [117797271, 117797302] |        +
3        chr1 [113171797, 113171828] |        +
4        chr1 [229997262, 229997293] |        -
5        chr1 [108697607, 108697638] |        +
6        chr1 [232944304, 232944335] |        +
```

The experiment has 7 samples. Following the workflow in Section 1.3, we start by removing over-amplification artifacts with `filterDuplReads`. We set the FDR at 0.01 and use 2 cores to speed up processing, and report the number of reads before and after filtering.

```
> gse25836Fil <- filterDuplReads(gse25836, fdrOverAmp=0.01, mc.cores=2)
> sapply(gse25836,nrow)
GSM634613 GSM634614 GSM634615 GSM634616 GSM634617 GSM634618 GSM634619
  3081174   5772530   3049264   5623018   2365943   4262704   2243477
> sapply(gse25836Fil,nrow)
GSM634613 GSM634614 GSM634615 GSM634616 GSM634617 GSM634618 GSM634619
  3079527   5763439   3047288   5615025   2365943   4261657   2241978
```

For illustration, we show the operations applied by `filterDuplReads` step-by-step. First we count the number of read repetitions, find the 0.999 quantile of the read repeat distribution and estimate the FDR for a series of cutoffs.

```
> tdr <- tabDuplReads(gse25836,mc.cores=2)
> head(tdr[['GSM634613']])
ans
      1       2       3       4       5       6
2488473  257510   22002    1868     261     101
> q <- sapply(tdr,function(z) which(cumsum(z/sum(z))>.999)[1])
> q
GSM634613.3 GSM634614.4 GSM634615.3 GSM634616.4 GSM634617.2 GSM634618.2
          3           4           3           4           2           2
GSM634619.3
          3
> fdrest <- vector("list",length(tdr)); names(fdrest) <- names(tdr)
> for (i in 1:length(fdrest)) fdrest[[i]] <- fdrEnrichedCounts(tdr[[i]],
use=1:q[i], components=0, mc.cores=2)
> fdrest[[1]][1:7,]
         pdfH0   pdfOverall fdrEnriched
1 8.984324e-01 8.982263e-01 1.000000000
2 9.297078e-02 9.294947e-02 0.997974636
3 7.943553e-03 7.941727e-03 0.974225969
4 6.070586e-04 6.742636e-04 0.740222673
5 4.310700e-05 9.420920e-05 0.221891062
6 2.905485e-06 3.645643e-05 0.027234825
7 1.883469e-07 2.129633e-05 0.002589592
> cutoff <- sapply(fdrest,function(z) rownames(z)[z$fdrEnriched<.01][1])
> cutoff
GSM634613 GSM634614 GSM634615 GSM634616 GSM634617 GSM634618 GSM634619
      "7"       "7"       "7"       "8"       "6"       "4"       "5"
```

The first sample has 2,488,473 reads appearing only once, 257,510 appearing twice, etc. Its 0.999 quantile for the read repeat distribution is $q_\alpha = 3$.

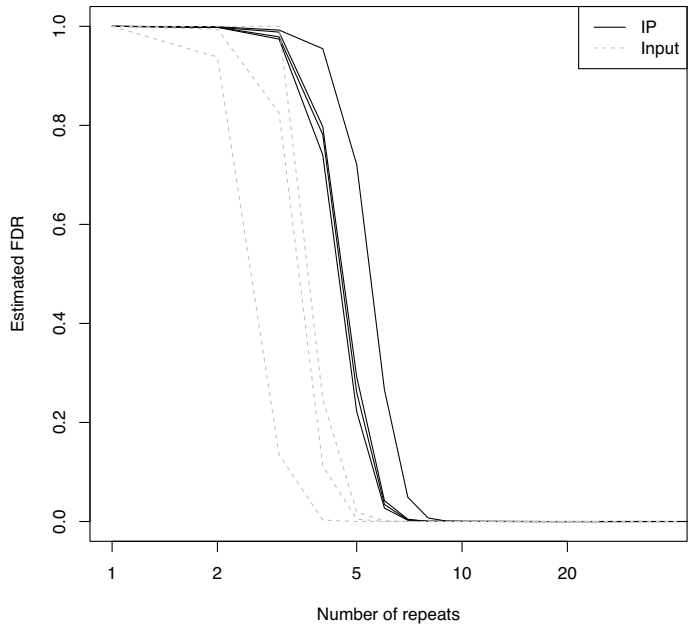


Figure 1: Estimated FDR vs. number of read repeats for GSE25836

Further, for this sample the removal of reads with ≥ 7 repeats is estimated to have an FDR≤ 0.0026.

Figure 1 shows the estimated FDR as a function of the number of read repeats for all samples. It is important to note that read repeats are more frequent in IP samples, which is to be expected as they focus on a relatively small part of the genome. That is, using the same cutoff for control and IP samples might be too stringent and result in a loss of precision in the subsequent peak calling. This observation illustrates the advantages of using a data-adaptive cutoff. The code required to produce Figure 1 is shown below.

```
> lty <- rep(1:2,c(4,3))
> col <- rep(c(1,'gray'),c(4,3))
> plot(fdrest[[1]]$fdrEnriched,type='l',xlab='Number of repeats',
ylab='Estimated FDR',log='x')
> for (i in 2:length(fdrest)) lines(fdrest[[i]]$fdrEnriched,lty=lty[i],
col=col[i])
> legend('topright',c('IP','Input'),lty=1:2,col=c(1,'gray'))
```

The next step in the workflow is to assess the similarity between samples via correlation plots. This is achieved with `cmds`, which also allows for parallel computing. Pearson, Spearman and Kendall correlation coefficients are available. In principle, Spearman is more general as it captures non-linear associations, but in practice all options typically produce very similar results.

```
> cmds1 <- cmds(gse25836Fil, mc.cores=2)
Computing coverage...
Computing correlations...
> cmds1Spear <- cmds(gse25836Fil, cor.method='spearman', mc.cores=2)
Computing coverage...
Computing correlations...
> dpearson <- cmds1@d[upper.tri(cmds1@d)]
> dspear <- cmds1Spear@d[upper.tri(cmds1Spear@d)]
> cor(dpearson,dspear)
[1] 0.9999551
> col <- rep(c(1,'gray'),c(4,3))
> pch <- c(1,2,1,2,1,2,1)
> n <- c('IP,GAI','IP,GAII','IP,GAI','IP,GAII','Input,GAI','Input,GAII',
'Input,GAI')
> plot(cmds1,col=col,pch=pch,labels=n)
```

The visual representation of the correlation distances can be achieved via classical Multi-Dimensional Scaling (MDS) and is shown in Figure 2(a). For comparison, Figure 2(b) shows a hierarchical clustering (complete linkage) dendrogram based on the SeqMonk correlations, which use RPKM in 1,000bp windows (necessary to avoid memory saturation problems in our 8Gb RAM desktop). Although the main features in the two plots are the same, panel (a) reveals some finer details, *e.g.* the relative position of GAI IP and input samples with respect to the GAII samples.

In order to remove potential strand-specific biases we use `alignPeaks`, which implements a procedure similar to that used by MACS (Zhang *et al*, 2008). Again, we make use of parallel computing by setting the argument `mc.cores`. A suitable alternative is provided in the package `chipseq` function `estimate.mean.fraglen`.

```
> gse25836 <- alignPeaks(gse25836Fil, strand='strand', mc.cores=4)
Estimated shift size is 7.389731
Estimated shift size is 7.99554
Estimated shift size is 0
Estimated shift size is 18.33544
Estimated shift size is 6.199713
Estimated shift size is 0.3745177
Estimated shift size is 0.9554126
```

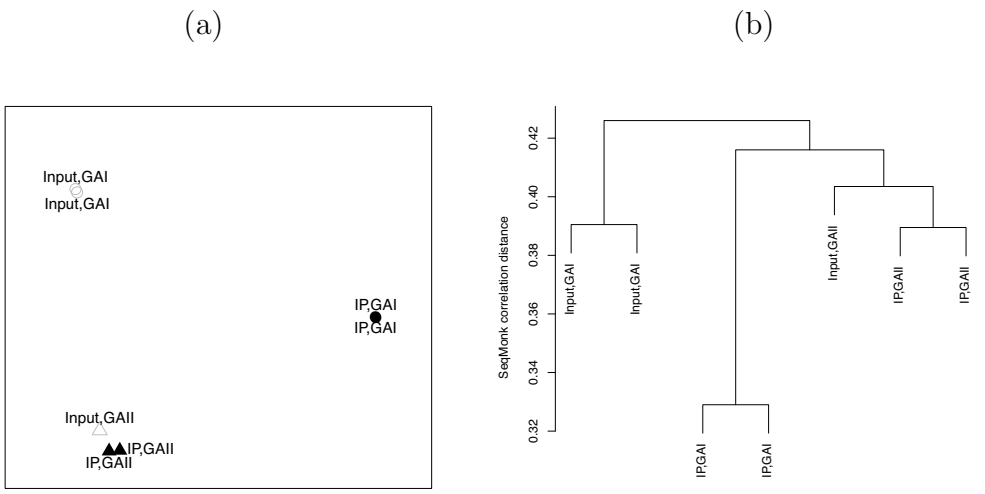


Figure 2: Sample correlations in GSE25836. (a) `htSeqTools` MDS plot; (b) SeqMonk hierarchical clustering (1,000bp window RPKM correlation)

In these data most samples require a minor adjustment only. Note that the adjustment tends to be slightly larger for the IP samples (the first 4 samples). In our experience, it is often the case that the stronger peaks observed in IP samples are more prone to strand-specific biases. See Section 3.2 or the `htSeqTools` vignette for examples where these biases are more evident.

Next we assess the efficiency of the immuno-precipitation with the n-adjusted coverage variability measures SD_n and G_n (Section 2.2). Both indexes are larger in the IP samples than in their respective controls, which suggests an efficient immuno-precipitation.

```
> sdn <- ssdCoverage(gse25836, mc.cores=6)
> gn <- giniCoverage(gse25836, mc.cores=6, mk.plot=FALSE)
> ans <- data.frame(n,round(cbind(sdn,gn),3))
> ans[c(1,3,5,7,2,4,6),]
                  n   sdn  gini gini.adjust
GSM634613     IP,GAI 0.115 0.975       0.006
GSM634615     IP,GAI 0.114 0.975       0.006
GSM634617  Input,GAI 0.110 0.978       0.002
GSM634619  Input,GAI 0.106 0.979       0.002
GSM634614    IP,GAII 0.124 0.951       0.014
GSM634616    IP,GAII 0.135 0.960       0.021
GSM634618 Input,GAII 0.120 0.955       0.002
```

After assessing that the data quality is satisfactory, the next step is to find peaks, *i.e.* chromosomal regions with large concentration of IP reads with respect to the input sample. To this end, we merge reads from all IP samples into a single `RangedData`, and similarly for the input samples. We use `enrichedRegions` to find regions where the coverage is greater than `minReads=10` (the default) and the proportion of reads in the IP sample falling in the region is significantly higher than in the input sample (Benjamini-Yekutieli adjusted P-value <0.05, see Section 2.3).

```
> ip <- rbind(gse25836[[1]],gse25836[[2]],gse25836[[3]],gse25836[[4]])
> input <- rbind(gse25836[[5]],gse25836[[6]],gse25836[[7]])
> regions <- enrichedRegions(ip,input,minReads=10,pvalFilter=.05,
p.adjust.method='BY',mc.cores=2)
> nrow(regions)
[1] 493
> head(regions)
RangedData with 6 rows and 5 value columns across 25 spaces
        space             ranges |   sample1   sample2       pvalue
  <character>          <IRanges> | <integer> <integer>    <numeric>
1        chr1 [7634643, 7634691] |        30         0 0.0037205770
2        chr1 [8230352, 8230440] |        44         0 0.0000907842
3        chr1 [8319253, 8319289] |        21         0 0.0369954205
4        chr1 [8319303, 8319359] |        32         0 0.0023097916
5        chr1 [8319784, 8319877] |        52         1 0.0002467893
6        chr1 [8323426, 8323483] |        31         1 0.0369954205
      rpkm1     rpkm2
  <numeric>  <numeric>
1  34.97487   0.000000
2  28.24188   0.000000
3  32.42265   0.000000
4  32.07053   0.000000
5  31.60141   1.199414
6  30.53266   1.943878
```

We find 493 enriched regions. For each region, the function reports the number of reads, RPKM (Mortazavi *et al*, 2008), and P-value. We visually inspect the first four detected peaks (Figure 3). The coverage in IP samples (black line) is above that for the controls (grey line) in all selected regions, as expected.

```
> par(mfrow=c(2,2))
> cov1 <- coverage(ip['chr1']); cov2 <- coverage(input['chr1'])
> for (i in 1:4) {
>   st <- start(regions['chr1'])[i]; en <- end(regions['chr1'])[i]
>   xlim <- c(st-100,en+100)
>   cov1sel <- seqselect(cov1[['chr1']],xlim[1],xlim[2])
>   cov2sel <- seqselect(cov2[['chr1']],xlim[1],xlim[2])
```

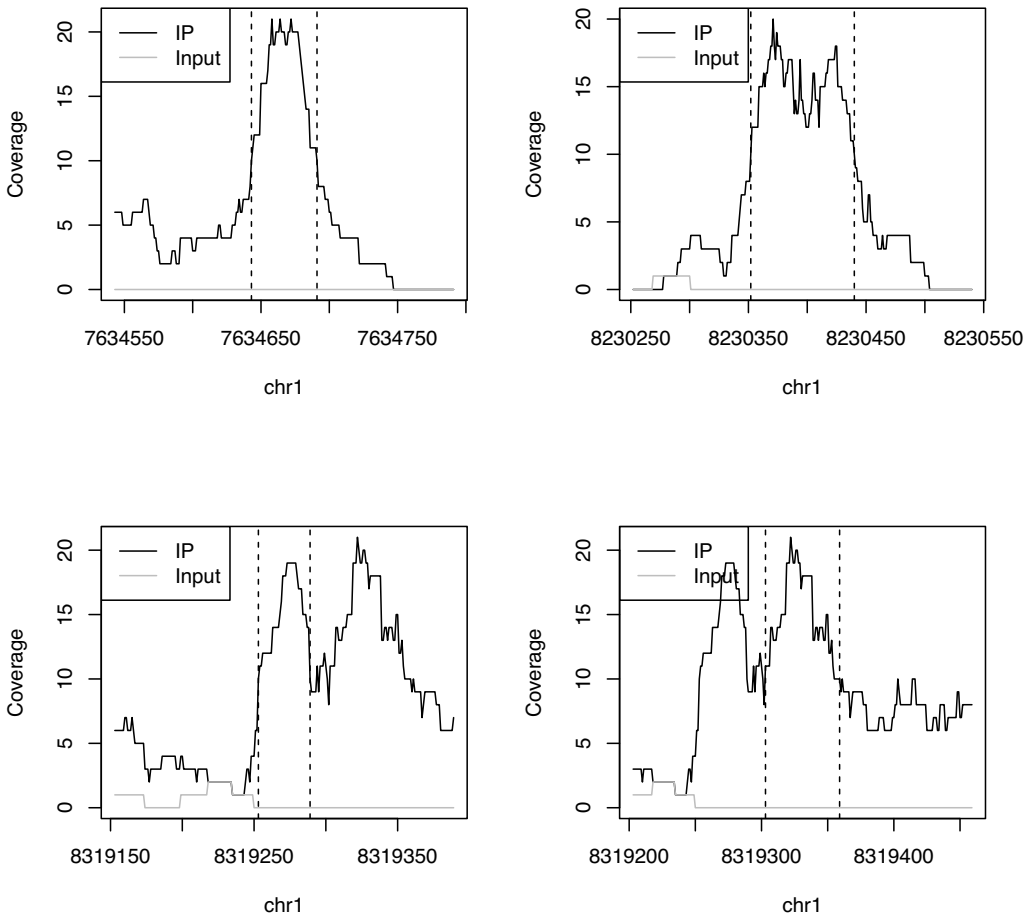


Figure 3: Four regions (±100bp) enriched in IP samples for the GSE25836 dataset. The vertical dashed lines indicate the enriched region limits. The lower panels show two detected peaks that were next to each other.

```
>   plot(xlim[1]:xlim[2],cov1sel,type='l',ylim=c(0,max(cov1sel)),
xlab='chr1',ylab='Coverage')
>   lines(xlim[1]:xlim[2],cov2sel,col='gray')
>   abline(v=c(st,en),lty=2)
>   legend('topleft',c('IP','Input'),col=c(1,'gray'),lty=c(1,1))
> }
```

Given that GAI and GAII data appear so separated in the MDS plot (Figure 2), another sensible strategy is to analyze GAI and GAII samples separately and report common findings.

```
> ip.ga1 <- rbind(gse25836[[1]],gse25836[[3]])
> input.ga1 <- rbind(gse25836[[5]],gse25836[[7]])
> ip.ga2 <- rbind(gse25836[[2]],gse25836[[4]])
> input.ga2 <- gse25836[[6]]
> regions.ga1 <- enrichedRegions(ip.ga1,input.ga1,minReads=10,
pvalFilter=.05,p.adjust.method='BY',mc.cores=2)
> regions.ga2 <- enrichedRegions(ip.ga2,input.ga2,minReads=10,
pvalFilter=.05,p.adjust.method='BY',mc.cores=2)
> nrow(regions.ga1)
[1] 267
> nrow(regions.ga2)
[1] 199
> sum(table(regions.ga1 %in% regions)[,2])
[1] 180
> sum(table(regions.ga2 %in% regions)[,2])
[1] 183
> sum(table(regions.ga1 %in% regions.ga2)[,2])
[1] 61
```

Most of the regions identified with the GAI and GAII samples individually are also found in the combined samples. There are 61 regions identified in both platforms. In order to keep the illustration simple, here we proceed with the results obtained from the merged samples.

Two useful functions related to `enrichedRegions` are `enrichedPeaks` and `islandCounts`. The former selects the highest coverage areas within the enriched regions, which can be useful to define peaks according to more stringent criteria. Below we define peaks as sub-regions with coverage above 25.

```
> peaks <- enrichedPeaks(regions,ip,input,minHeight=25,mc.cores=2)
> nrow(peaks)
[1] 314
> head(peaks)
RangedData with 6 rows and 2 value columns across 25 spaces
        space               ranges |    height region.pvalue
  <character>            <IRanges> | <integer>     <numeric>
```

```
1        chrY [13458415, 13458431] |        53  1.032031e-05
2        chrY [13458468, 13458478] |        35  1.032031e-05
3        chrY [13460873, 13460874] |        26  1.480397e-02
4        chrY [13468081, 13468086] |        31  5.249346e-05
5        chrY [13468126, 13468133] |        39  5.249346e-05
6        chrY [13468135, 13468138] |        28  5.249346e-05
```

`islandCounts` finds all regions with overall coverage (*i.e.* across all samples) above a user-specified threshold. Next it computes the number of reads overlapping with each region. This function is useful to obtain read counts, which can then be analyzed with a number of Bioconductor packages, *e.g.* `BayesPeak` (Spyrou *et al*, 2009), `DESeq` (Anders and Huber, 2010) and `edgeR` (Robinson *et al*, 2010). Here we show how to combine `islandCounts` with package `edgeR`.

```
> counts <- islandCounts(RangedDataList(ip,input),minReads=10,mc.cores=2)
> head(counts)
RangedData with 6 rows and 2 value columns across 25 spaces
        space             ranges |   counts1   counts2
  <character>          <IRanges> | <integer> <integer>
1        chr1 [1336314, 1336315] |         9         1
2        chr1 [1336318, 1336329] |        10         1
3        chr1 [2507532, 2507532] |        10         0
4        chr1 [2507549, 2507551] |        10         0
5        chr1 [5732170, 5732193] |         7         6
6        chr1 [5734147, 5734172] |        13         3
> nrow(counts)
[1] 6572
> countsTable <- cbind(counts[['counts1']],counts[['counts2']])
> library(edgeR)
> d <- DGEList(countsTable,lib.size=c(nrow(ip),nrow(input)),
group=c('ip','input'))
> d <- estimateCommonDisp(d)
Warning message:
In estimateCommonDisp(d) :
  There is no replication.  Setting common dispersion to 0.
> de.com <- exactTest(d)
Comparison of groups:  ip - input
> padj <- p.adjust(de.com$table$p.value,method='BY')
> sel <-  (padj<.05) & (de.com$table$logFC>0)
> sum(sel)
[1] 413
> tab <- table(counts[sel,] %in% regions)
> sum(tab[,2])
[1] 380
```

`islandCounts` finds 6572 regions. The exact test in `edgeR` determined that 413 are significantly enriched in the IP samples. Of these, 380 (92%)

overlap with the regions found by `enrichedRegions`. Since no technical replicates are available, `edgeR` estimates the variability based on a Poisson model. Below we illustrate how to obtain more precise variance estimates by using the technical replicates obtained with the GAII platform. Note that due to having a smaller number of reads, in this analysis we choose not to adjust P-values in order to preserve statistical power. We identify 61 regions, all of which were also detected in our previous analysis.

```
> counts.ga2 <- islandCounts(gse25836[c(2,4,6)],minReads=10,mc.cores=4)
> lsize <- sapply(gse25836,nrow)[c(2,4,6)]
> countsTable.ga2 <- cbind(counts.ga2[['GSM634614']],
counts.ga2[['GSM634616']],counts.ga2[['GSM634618']])
> d <- DGEList(countsTable.ga2,lib.size=lsize,group=c('ip','ip','input'))
> d <- estimateCommonDisp(d)
> de.com.ga2 <- exactTest(d)
Comparison of groups:  ip - input
> sel <-  (de.com.ga2$table$p.value<.05) & (de.com.ga2$table$logFC>0)
> tab <- table(counts.ga2[sel,] %in% regions)
> sum(tab['TRUE'])
[1] 61
> sum(tab['TRUE'])/sum(tab)
[1] 1
```

We show how to use packages `biomaRt` (Durinck and Huber, 2011) and `ChIPpeakAnno` to annotate the regions by finding the closest transcription start site (TSS). The peak distribution with respect to the closest TSS can be visualized with `PeakLocation` and `stdPeakLocation`. Figure 4(a) shows that although most peaks occur close to the TSS, a non-negligible proportion locate further downstream. Since transcripts have different lengths, panel (a) does not reveal their exact location. The distance relative to the transcript length in panel (b) reveals that most downstream peaks locate close to the transcript end.

```
> library(biomaRt)
> library(ChIPpeakAnno)
> mart <- useMart("ensembl", "hsapiens_gene_ensembl")
> hsanno <- getAnnotation(mart, featureType='TSS')
> peaksanno <- annotatePeakInBatch(peaks, AnnotationData=hsanno,
PeakLocForDistance="middle")
> PeakLocation(peaksanno,peakDistance=10^4)
> stdPeakLocation(peaksanno)
```

As a final step in our proposed ChIP-seq workflow, we look for chromosome regions with a large number of peaks.

We indicate the chromosome lengths in a named vector, use `nSims=100` simulations to estimate the FDR and set the `mc.cores` argument to speed up

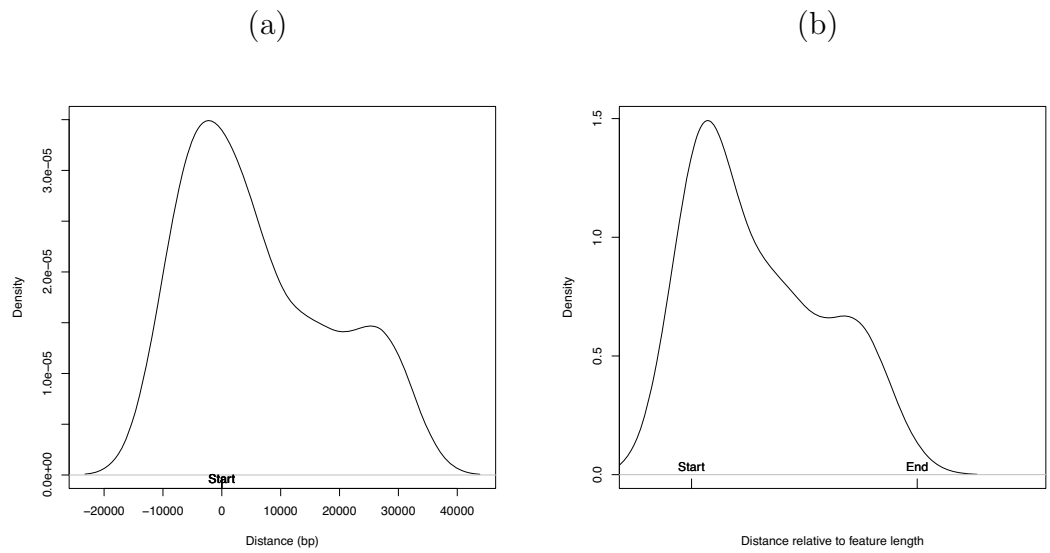


Figure 4: Distribution of distances to closest TSS. (a) Absolute distance in bp; (b) Distance relative to transcript length

computations. We find 32 chromosomal areas where peaks tend to cluster. Figure 5 shows their location.

```
> library(BSgenome.Hsapiens.UCSC.hg19)
> chrLength <- seqlengths(Hsapiens)[names(ip)]
> chrRegions <- enrichedChrRegions(regions,chrLength=chrLength,fdr=0.05,
nSims=100,mc.cores=6)
> nrow(chrRegions)
[1] 32
> plotChrRegions(chrRegions,chrLength=chrLength)
```

3.2 Example 2: GSE16926

Here we illustrate the quality control features in `htSeqTools` with the *S. cerevisiae* ChIP-seq data GSE16926 (www.ncbi.nlm.nih.gov/geo). We read the data into R via `read.table` and stored it into a `RangedDataList` object `gse16926Raw`. We load `htSeqTools`, `multicore` and the data. We remove over-amplification artifacts with `filterDuplReads` (using 6 cores to speed up computations).

```
> library(htSeqTools)
> library(multicore)
```

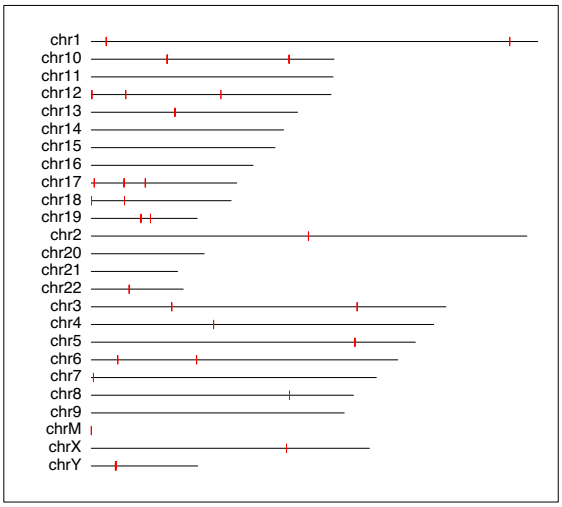


Figure 5: Chromosome areas with high peak density in GSE25836

```
> load('gse16926Raw.RData')
> gse16926Raw
RangedDataList of length 8
names(8): GSM424491 GSM424492.rep1 GSM424492.rep2 ... GSM424494.rep2
GSM442537
> gse16926 <- filterDuplReads(gse16926Raw,mc.cores=6)
```

Similar to the example in Section 3.1, we show the estimated FDR for each possible threshold to declare over-amplification, *i.e.* maximum number of allowable repeats (Figure 6).

```
> tdr <- tabDuplReads(gse16926Raw,mc.cores=6)
> q <- sapply(tdr,function(z) which(cumsum(z/sum(z))>.999)[1])
> q
     GSM424491.35 GSM424492.rep1.47 GSM424492.rep2.55 GSM424493.rep1.38
               35                 47                 55                 38
GSM424493.rep2.35 GSM424494.rep1.91 GSM424494.rep2.36      GSM442537.40
               35                 91                 36                 40
> fdrest <- vector("list",length(tdr)); names(fdrest) <- names(tdr)
> for (i in 1:length(fdrest)) fdrest[[i]] <- fdrEnrichedCounts(tdr[[i]],
use=1:q[i], components=0, mc.cores=6)
>
```

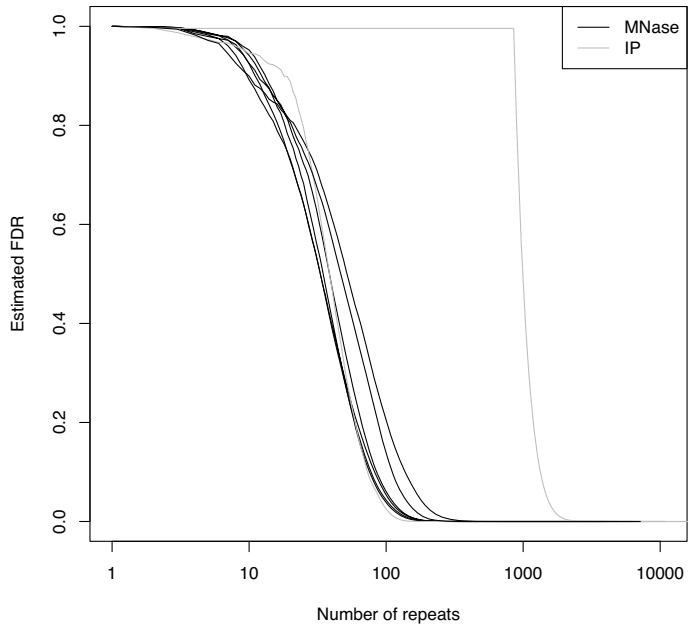


Figure 6: Estimated FDR vs. number of read repeats for GSE16926

```
> col <- c(1,1,1,1,1,'grey','grey',1)
> plot(fdrest[[1]]$fdrEnriched,type='l',xlab='Number of repeats',
ylab='Estimated FDR',log='x')
> for (i in 2:length(fdrest)) lines(fdrest[[i]]$fdrEnriched,
col=col[i])
> legend('topright',c('MNase','IP'),lty=1,col=c(1,'gray'))
```

The first replicate of sample GSM424494 shows an unusually large number of highly repetitive reads (0.999 quantile= 91 repeats). The sample also stands out in Figure 6. The observation that GSM424494 was immunoprecipitated whereas the rest of the samples were MNase-digested does not explain this finding, as this behavior is not observed in the second replicate. This finding suggests that the first replicate was particularly prone to over-amplification artifacts. By default, `filterDuplReads` adopts a conservative approach and allows for a larger number of read repetitions in this sample. A more aggressive repeat removal can be forced by setting the argument `maxRepeats`.

```
> gsm424494 <- filterDuplReads(gse16926Raw[[6]],maxRepeats=500,mc.cores=6)
```

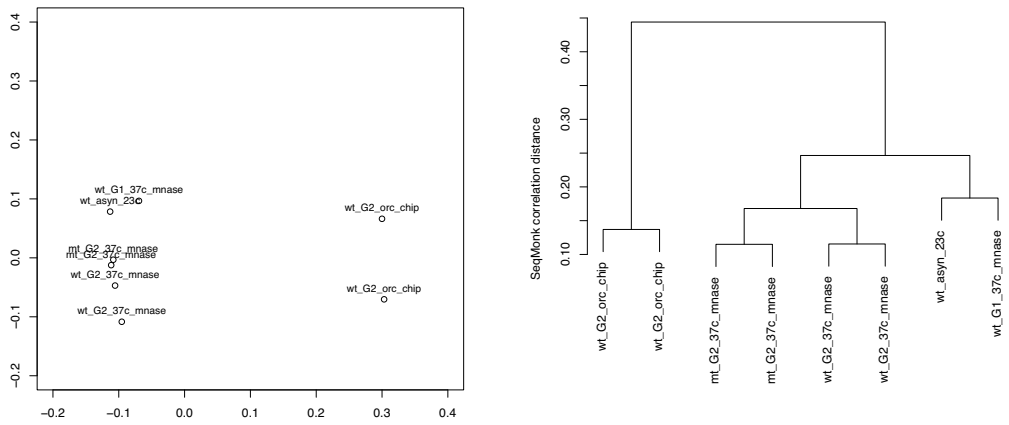


Figure 7: Sample correlationsin GSE16926. (a) `htSeqTools` MDS plot; (b) SeqMonk hierarchical clustering

We produce an MDS plot to visualize the correlations between samples (Figure 7(a)). The plot reveals several interesting features. First, IP and MNase-digested samples are clearly separated. Second, the mutant G2 samples are most similar to the wild-type G2 samples. In fact, mutant G2 samples present coverage profiles which are somewhere in between wild-type G2 and wild-type G1 and asynchronous samples. The latter observation is not obvious from the dendrogram in Figure 7(b), which is based on correlations computed with the SeqMonk software. This example illustrates the potential advantages in visualizing similarity (or dissimilarity) matrices via MDS plots.

```
> lab <- c("wt_asyn_23c","wt_G2_37c_mnase","wt_G2_37c_mnase",
"mt_G2_37c_mnase", "mt_G2_37c_mnase","wt_G2_orc_chip",
"wt_G2_orc_chip","wt_G1_37c_mnase")
> plot(cmds1,labels=lab,cex.text=.8,xlim=c(-.2,.4))
```

Next we use `alignPeaks` to remove strand-specific biases and assess the sample enrichment efficiency via the indexes SD_n and G_n (Section 2.2).

```
> gse16926 <- alignPeaks(gse16926, strand='strand', mc.cores=6)
Estimated shift size is 10.71125
Estimated shift size is 32.54875
Estimated shift size is 39.20407
Estimated shift size is 22.65083
```

```
Estimated shift size is 30.42685
Estimated shift size is 41.10548
Estimated shift size is 45.36787
Estimated shift size is 18.31953
> sdn <- ssdCoverage(gse16926, mc.cores=6)
> gn <- giniCoverage(gse16926, mc.cores=6, mk.plot=FALSE)
> data.frame(lab,round(cbind(sdn,gn),3))
                           lab    sdn  gini gini.adjust
GSM424491          wt_asyn_23c  9.300 0.499       0.360
GSM424492.rep1 wt_G2_37c_mnase  9.388 0.512       0.362
GSM424492.rep2 wt_G2_37c_mnase  9.768 0.530       0.353
GSM424493.rep1 mt_G2_37c_mnase  9.855 0.540       0.396
GSM424493.rep2 mt_G2_37c_mnase  8.378 0.520       0.349
GSM424494.rep1  wt_G2_orc_chip 24.163 0.655       0.458
GSM424494.rep2  wt_G2_orc_chip 13.661 0.594       0.346
GSM442537      wt_G1_37c_mnase  9.604 0.490       0.317
```

Most samples require a strand-bias correction ranging between 20-50 bases, which is non-negligible. Both SD_n and G_n suggest that the IP sample GSM424494 (replicate 1) presents the clearest peaks. According to SD_n, GSM424494 replicate 2 is the second sample presenting clearest peaks. These findings suggest that immuno-precipitation of sonicated DNA was more efficient than MNase digestion in isolating the target DNA. This would be consistent with its capacity to produce shorter DNA fragments. It should also be noted that SD_n gives more consistent results across replicates than G_n, thereby suggesting SD_n as the default metric of choice.

3.3 Example 3: Histone methylation data

In Section 3.1 and 3.2 we showed that MDS correlation plots and the indexes SD_n and G_n can help assess inefficient immuno-precipitation of DNA-binding transcription factors. We now explore the case of histone methylation marks, which in some cases may show less pronounced coverage profiles than transcription factors and therefore represent a more challenging scenario.

We obtained ChIP-seq data assessing the genome-wide distribution of histones H3K9Me3 and H3K4Me3 during the *D. melanogaster* development. For H3K9Me3, aligned reads in BAM format were downloaded from the modEncode project (modEncode 801-809) and imported into R using `scanBAM` from package `Rsamtools` (Morgan and Pagés, 2011a). The H3K4Me3 raw data were obtained in FASTQ format from GEO (GSE15292). Reads uniquely aligned with Bowtie to the dm3 genome with up to two mismatches were read into R with the `readAligned` function in package `ShortRead`.

Figure 8 shows an MDS plot to visualize the Pearson correlation between log-coverages for H3K9Me3. Except in Pupae, the plot shows a large separa-

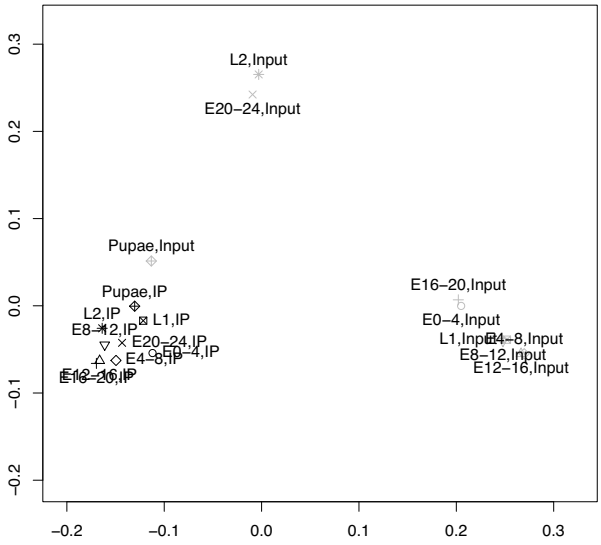


Figure 8: MDS plot for coverage correlation in H3K9Me3 data (modEncode 801-809).

	H3K9Me3		H3K4Me3	
	SD_n	G_n	SD_n	G_n
E0-4, Input	0.631	0.079	0.640	0.098
E0-4, IP	0.749	0.131	1.157	0.252
E4-8, Input	0.636	0.195	0.638	0.225
E4-8, IP	0.712	0.147	2.169	0.380
E8-12, Input	0.520	0.112	0.516	0.138
E8-12, IP	0.714	0.190	2.505	0.478
E12-16, Input	0.774	0.226	0.824	0.265
E12-16, IP	0.714	0.176	2.520	0.427
E16-20, Input	0.612	0.115	0.620	0.140
E16-20, IP	0.764	0.200	1.825	0.411
E20-24, Input	0.511	0.043	0.509	0.043
E20-24, IP	0.578	0.078	1.255	0.337
L1, Input	0.598	0.138	0.604	0.169
L1, IP	0.508	0.043	0.636	0.179
L2, Input	0.489	0.033	0.483	0.033
L2, IP	0.704	0.211	1.219	0.345
L3, Input	-	-	0.501	0.084
L3, IP	-	-	0.858	0.250
Pupae, Input	0.514	0.130	0.510	0.130
Pupae, IP	0.537	0.089	0.602	0.133
Adult female, Input	-	-	0.615	0.0734
Adult female, IP	-	-	1.257	0.300
Adult male, Input	-	-	0.509	0.121
Adult male, IP	-	-	0.513	0.072

Table 3: SD_n and G_n for H3K9Me3 (modEncode 801-809) and H3K4Me3 (GSE15292) ChIP-seq data

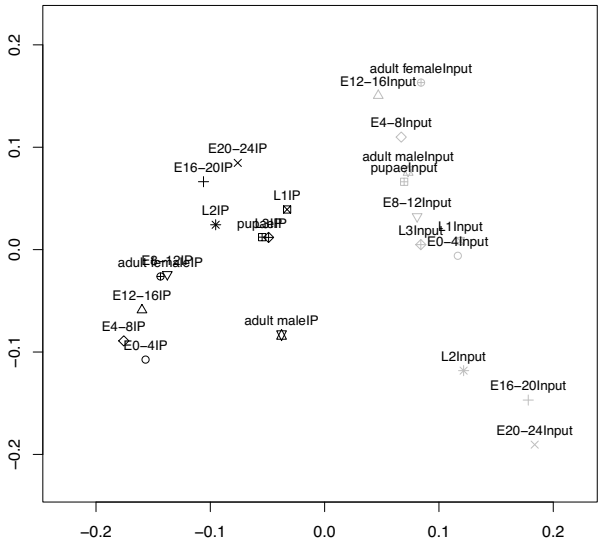


Figure 9: MDS plot for coverage correlation in H3K4Me3 data (GSE25836).

tion between IP and input samples. That is, the immuno-precipitation was efficient enough to give samples a distinctive coverage profile. The two samples with the lowest number of aligned reads are the E20-24 and L2 inputs, at 1.05 and 0.67 million reads (respectively). These two samples appear as outliers in Figure 8. Regarding H3K4Me3, Figure 9 also shows a clear separation between IP and control samples.

Table 3 shows SD_n and G_n for the H3K9Me3 and H3K4Me3 data. In most developmental stages, the H3K9Me3 IP samples present a larger SD_n than their controls. The IP Pupae sample presents only a slightly larger SD_n than its control, which is consistent with Figure 8 in signaling poor sample separation. The poor separation might be due to the sequencing depth in the Pupae samples being fairly low. SD_n and G_n are designed not to be biased by the sequencing depth, but their variability can be relatively large when few reads are available. In this example, the MDS plot also suggests poor separation and we therefore deem it likely that there were some problems in the Pupae sample preparation. For embryo 12-16 hours (E12-16) and Larva 1 (L1), SD_n is larger in the controls. Figure 8 shows a clear separation between E12-16/L1 IP and control samples. SD_n is useful in signaling that downstream peak calling algorithms may need to be adjusted

to take into consideration that E12-16/L1 peaks are not as pronounced as in the other samples. The index G_n does not discriminate between IP and input samples as well as SD_n. Similar to SD_n, G_n flags the Pupae, E12-16 and L1 samples, and the E4-8 sample as well. Regarding H3K4Me3, both SD_n and G_n indicate the presence of stronger peaks than for H3K9Me3. SD_n consistently presents larger values in the IP samples, which agrees with Figure 9 in signaling efficient immuno-precipitation. G_n points in the same direction for all samples, with the exception of E4-8 and the adult male.

All together, these results suggest that the combined use of MDS and our proposed indexes can be a useful diagnosis for histone methylation data, signaling either problems in the sample preparation or the absence of pronounced peaks. Also, SD_n appears to be more sensitive than G_n, and we therefore recommend it as a default choice.

3.4 Example 4: Yeast RNA-seq

We illustrate the RNA-seq workflow (Section 1.4) with the Yeast RNA-seq data in the Bioconductor package `yeastRNASeq` (Lee *et al*, 2010). Following the `yeastRNASeq` vignette, we import the Bowtie aligned reads and format the data as a `RangedDataList`.

```
> library(yeastRNASeq)
> library(ShortRead)
> f <- list.files(file.path(system.file(package = "yeastRNASeq"),"reads"),
pattern="bowtie", full.names=TRUE)
> names(f) <- gsub("\\.bowtie.*", "", basename(f))
> names(f)
[1] "mut_1_f" "mut_2_f" "wt_1_f"  "wt_2_f"
> aligned <- lapply(f, readAligned, type = "Bowtie")
> aligned
$mut_1_f
class: AlignedRead
length: 423318 reads; width: 26 cycles
chromosome: Scchr05 Scchr15 ... Scchr08 Scchr13
position: 541317 885627 ... 488228 667296
strand: - + ... - +
alignQuality: NumericQuality
alignData varLabels: similar mismatch

$mut_2_f
class: AlignedRead
length: 420848 reads; width: 26 cycles
chromosome: Scchr02 Scchr13 ... Scchr16 Scchr04
position: 787251 120582 ... 719121 790753
strand: + + ... - -
```

```
alignQuality: NumericQuality
alignData varLabels: similar mismatch

$wt_1_f
class: AlignedRead
length: 410349 reads; width: 26 cycles
chromosome: Scchr11 Scchr04 ... Scchr05 Scchr10
position: 283246 961713 ... 195595 218358
strand: + - ... + +
alignQuality: NumericQuality
alignData varLabels: similar mismatch

$wt_2_f
class: AlignedRead
length: 430264 reads; width: 26 cycles
chromosome: Scchr14 Scchr04 ... Scchr13 Scchr06
position: 705342 1283088 ... 17545 254886
strand: - - ... - +
alignQuality: NumericQuality
alignData varLabels: similar mismatch

> al2rd <- function(z) RangedData(IRanges(position(z),position(z)+26),
space=chromosome(z),strand=strand(z))
> seqs <- lapply(aligned,al2rd)
> seqs <- RangedDataList(lapply(aligned,al2rd))
> seqs
RangedDataList of length 4
names(4): mut_1_f mut_2_f wt_1_f wt_2_f
```

We remove over-amplification artifacts with `filterDuplReads`, and display the estimated FDR for each threshold on the maximum number of allowable repeats (Figure 10(a)). The estimated FDR is similar across all samples. The MDS coverage correlation plot (Figure 10(b)) reveals a clear separation between wild-type and mutant samples.

```
> seqsFilt <- filterDuplReads(seqs,mc.cores=4)
> sapply(seqs,nrow)
mut_1_f mut_2_f  wt_1_f  wt_2_f
 423318  420848  410349  430264
> sapply(seqsFilt,nrow)
mut_1_f mut_2_f  wt_1_f  wt_2_f
 418885  416768  409773  429763
> tdr <- tabDuplReads(seqs,mc.cores=4)
> q <- sapply(tdr,function(z) which(cumsum(z/sum(z))>.999)[1])
> fdrest <- vector("list",length(tdr)); names(fdrest) <- names(tdr)
> for (i in 1:length(fdrest)) fdrest[[i]] <- fdrEnrichedCounts(tdr[[i]],
use=1:q[i], components=0, mc.cores=6)
> col <- c(1,1,'grey','grey')
```

(a) (b)

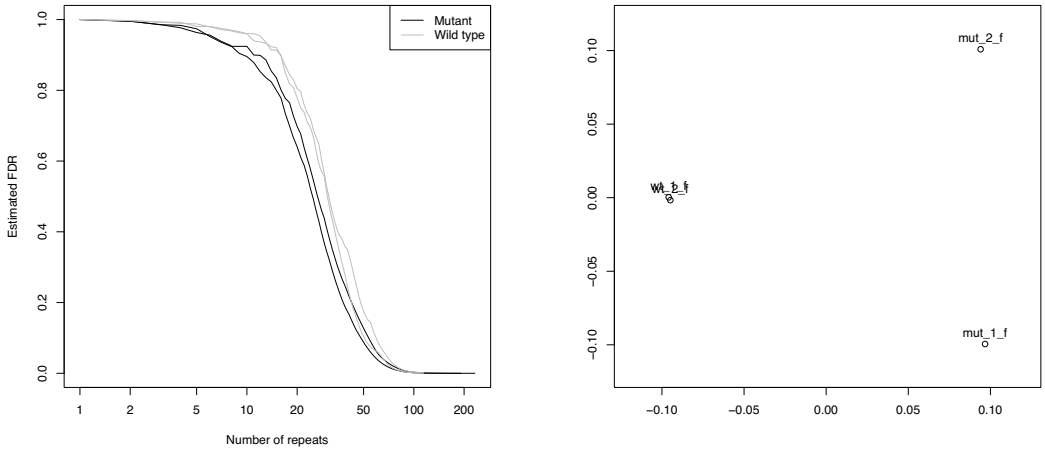


Figure 10: Quality control for Yeast RNA-seq data. (a) Estimated FDR vs. number of read repeats (b) MDS plot to visualize correlation between coverages

```
> plot(fdrest[[1]]$fdrEnriched,type='l',xlab='Number of repeats',
ylab='Estimated FDR',log='x')
> for (i in 2:length(fdrest)) lines(fdrest[[i]]$fdrEnriched,col=col[i])
> legend('topright',c('Mutant','Wild type'),lty=1,col=c(1,'gray'))
>
> mds1 <- cmds(seqsFilt,mc.cores=4)
Computing coverage...
Computing correlations...
> plot(mds1)
```

As the dataset focuses on novel low abundance and transient RNAs, the goal is to detect RNAs de novo and compare their expression between wild-type and mutant. We use `islandCounts` to find all genomic regions with $\geq$10 reads (across all samples) and compute the number of reads in each sample.

```
> islands1 <- islandCounts(seqs,minReads=10,mc.cores=4)
> nrow(islands1)
[1] 29934
> head(islands1)
RangedData with 6 rows and 4 value columns across 17 spaces
        space         ranges |   mut_1_f   mut_2_f    wt_1_f    wt_2_f
  <character>      <IRanges> | <integer> <integer> <integer> <integer>
```

```
1     Scchr01 [33451, 33453] |         2         0         4         4
2     Scchr01 [33456, 34741] |       104        95       754       776
3     Scchr01 [34745, 34765] |         2         1         5         6
4     Scchr01 [34785, 34935] |        17        20        71        73
5     Scchr01 [35018, 35023] |         3         0         5         6
6     Scchr01 [35026, 35058] |         3         1        14         8
```

As shown below, it is possible to compare the total read count in each group using the function `enrichedRegions`. We also illustrate how to analyze the count data with the `edgeR` package. In our opinion, `edgeR` is preferable whenever replicates are available, as it estimates the within-group variability. In this particular example both approaches give similar results. Indeed, all regions found to be statistically significant by `edgeR` are also detected by `enrichedRegions`.

```
> regions <- RangedData(ranges(islands1))
> regions$sample1 <- islands1$mut_1_f+islands1$mut_2_f
> regions$sample2 <- islands1$wt_1_f+islands1$wt_2_f
> regions <- enrichedRegions(regions=regions,pvalFilter=.05,
p.adjust.method='BY',twoTailed=TRUE)
> head(regions)
RangedData with 6 rows and 5 value columns across 17 spaces
        space         ranges |   sample1   sample2       pvalue
  <character>      <IRanges> | <integer> <integer>    <numeric>
1     Scchr01 [33456, 34741] |       199      1530 0.000000e+00
2     Scchr01 [34785, 34935] |        37       144 5.248181e-11
3     Scchr01 [35204, 35210] |         0        11 3.805142e-02
4     Scchr01 [35270, 35363] |        11        56 5.213262e-05
5     Scchr01 [35371, 35508] |        14       112 0.000000e+00
6     Scchr01 [35530, 35702] |        31       104 3.222074e-06
  rpkm.sample1 rpkm.sample2
     <numeric>    <numeric>
1    267.3110     1840.8981
2    423.2817     1475.5870
3      0.0000     2431.4981
4    202.1480      921.8058
5    175.2482     1255.7935
6    309.5425      930.1790
>
> library(edgeR)
> group <- c('mut','mut','wt','wt')
> d <- DGEList(countsTable,lib.size=colSums(countsTable),group=group)
> d <- estimateCommonDisp(d)
>
> de.com <- exactTest(d)
Comparison of groups:  wt - mut
> padj <- p.adjust(de.com$table$p.value,method='BY')
> deIslands <- islands1[padj<.05,]
```

```
> nrow(deIslands)
[1] 1133
>
> tab <- table(deIslands %in% regions)
> sum(tab[,'TRUE'])/sum(tab)
[1] 1
```

The next step in the workflow is to identify genomic features close to the differentially expressed RNAs. We perform this step with the `biomaRt` (Durinck and Huber, 2011) and `ChIPpeakAnno` (Zhu *et al*, 2011) packages. For illustration purposes, we find the closest transcription start site. This requires adjusting the chromosome names so that they match those provided by `biomaRt`.

```
> library(biomaRt)
> library(ChIPpeakAnno)
> mart <- useMart("ensembl", "scerevisiae_gene_ensembl")
> yeastanno <- getAnnotation(mart, featureType='TSS')
>
> newchr <- as.character(deIslands$space)
> newchr[newchr=='Scchr01'] <- 'I'
> newchr[newchr=='Scchr02'] <- 'II'
> newchr[newchr=='Scchr03'] <- 'III'
> newchr[newchr=='Scchr04'] <- 'IV'
> newchr[newchr=='Scchr05'] <- 'V'
> newchr[newchr=='Scchr06'] <- 'VI'
> newchr[newchr=='Scchr07'] <- 'VII'
> newchr[newchr=='Scchr08'] <- 'VIII'
> newchr[newchr=='Scchr09'] <- 'IX'
> newchr[newchr=='Scchr10'] <- 'X'
> newchr[newchr=='Scchr11'] <- 'XI'
> newchr[newchr=='Scchr12'] <- 'XII'
> newchr[newchr=='Scchr13'] <- 'XIII'
> newchr[newchr=='Scchr14'] <- 'XIV'
> newchr[newchr=='Scchr15'] <- 'XV'
> newchr[newchr=='Scchr16'] <- 'XVI'
> newchr[newchr=='Scmito'] <- 'Mito'
> deIslands <- RangedData(IRanges(start(deIslands),end(deIslands)),
values=values(deIslands),space=newchr)
> islandsAnno <- annotatePeakInBatch(deIslands, AnnotationData=yeastanno,
PeakLocForDistance="middle")
```

The final step in our basic workflow is to find chromosomal regions accumulating differentially expressed RNAs using `enrichedChrRegions`. We identify 62 significantly enriched regions (FDR$\leq$0.05), shown in Figure 11.

```
> library(BSgenome.Scerevisiae.UCSC.sacCer2)
```

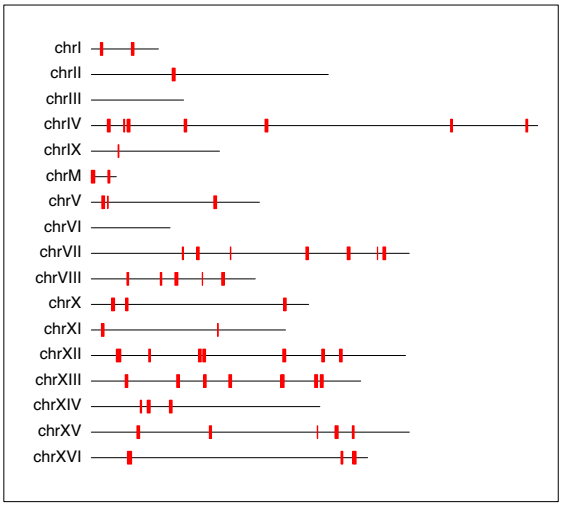


Figure 11: Yeast RNA-seq data: genomic regions with high concentration of differentially expressed RNAs

```
> newchr <- paste('chr',deIslands$space,sep='')
> newchr[newchr=='chrMito'] <- 'chrM'
> deIslands <- RangedData(IRanges(start(deIslands),end(deIslands)),
values=values(deIslands),space=newchr)
> chrLength <- seqlengths(Scerevisiae)[names(deIslands)]
> chrRegions <- enrichedChrRegions(deIslands,chrLength=chrLength,
fdr=0.05,nSims=100,mc.cores=6)
> nrow(chrRegions)
[1] 62
> plotChrRegions(chrRegions,chrLength=chrLength)
```

References

Anders,S., Huber,W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, **11**, R106.

Benjamini,Y., Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, **29**, pp. 1165-1188.

Durinck,S., Huber,W. biomaRt: Interface to BioMart databases (e.g. Ensembl, COSMIC, Wormbase and Gramene). R package version 2.6.0.

Efron,B., Tibshirani,R., Storey,J.D., Tusher,V. (2001) Empirical Bayes analysis of a microarray experiment, *Journal of the American Statistical Association*, **96**, 1151-1160.

Gentleman,R.C., Carey,V.J., Bates,D.M., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier, L., Ge,Y., Gentry,J., Hornik,K., Hothorn,T., Huber,W., Iacus,S., Irizarry,R., Leisch,F., Li,C., Maechler,M., Rossini,A.J., Sawitzki,G., Smith,C., Smyth,G., Tierney,L., Yang,J.Y.H., Zhang,J.. Bioconductor: Open software development for computational biology and bioinformatics (2004) *Genome Biology*, **5**:R80.

Gini,C. (1912) *Variabilita e mutabilita*. C. Cuppini, Bologna.

Karolchik,D., Hinrichs,A.S., Furey,T.S., Roskin,K.M., Sugnet,C.W., Haussler,D., Kent,W.J. (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Research*, **32**, pp. D493-496

Lee,A., Hansen,K.D., Bullard,J., Dudoit,S., Sherlock,G. (2008). Novel low abundance and transient RNAs in yeast revealed by tiling microarrays and ultra high-throughput sequencing are not conserved across closely related yeast species. *PLoS Genetics*, **4**(12): 41000299.

McLean,C.Y., Bristor,D., Hiller,M., Clarke,S.L., Schaar,B.T., Lowe,C.B., Wenger,A.M., Bejerano,G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology*, **28**(5), pp. 495-501.

Morgan,M., Lawrence,M., Anders,S. Classes and methods for high-throughput short-read sequencing data, R package version 1.10.4 (`http://www.bioconductor.org/packages/2.8/bioc/html/ShortRead.html`)

Morgan,M., Pagés,H. Rsamtools: Import aligned BAM file format sequences into R / Bioconductor, R package version 1.2.3 (`http://bioconductor.org/packages/release/bioc/html/Rsamtools.html`)

Mortazavi,A., Williams,B.A., McCue,K., Schaeffer,L., Wold,B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature methods*, **5**(7), pp. 621-628.

Lawrence,M., Carey,V., Gentleman,R. rtracklayer: R interface to genome browsers and their annotation tracks, R package version 1.10.6 (`http://www.bioconductor.org/packages/2.8/bioc/html/rtracklayer.html`)

Pagés,H., Aboyoun,P., Lawrence,M. IRanges: Infrastructure for manipulating intervals on sequences, R package version 1.8.9. (`http://www.bioconductor.org/packages/2.8/bioc/html/IRanges.html`)

Robinson,M.D., McCarthy,D.J., Smyth,G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, pp. 139-140.

Schwarz,G.E. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**(2), pp. 461464.

Spyrou,C., Stark,R., Lynch,A.G., Tavaré,S. (2009). BayesPeak: Bayesian analysis of ChIP-seq data. *BMC Bioinformatics*, **10**:299.

Urbanek,S. multicore: Parallel processing of R code on machines with multiple cores or CPUs, R package version 0.1-5 (`http://CRAN.R-project.org/package=multicore`)

Wang,L., Wang,X. (2011) DEGseq: Identify Differentially Expressed Genes from RNA-seq data. R package version 1.4.3.

Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nussbaum,C., Myers,R.M., Brown,M., Li,W., Liu,X.S. (2008) Model-based Analysis of ChIP-Seq (MACS), *Genome Biology*, **9**, R173.

Zhu,L.J., Pages,H., Gazin,C., Lawson,N., Lin,S., Lapointe,D., Green,M. (2009). ChIPpeakAnno: Batch annotation of the peaks identified from either ChIP-seq or ChIP-chip experiments. R package version 1.6.0.

8. DASiR: Programmatic data retrieval from DAS servers in R

Additional materials are available in the Bioconductor repository,

http://bioconductor.org/packages/release/bioc/html/DASiR.html,

including the source code and the reference manual