Nucleosome dynamics and analysis in breast cancer cells

Andy Pohl

Tesi Doctoral UPF / Any 2014

Director de la tesi

Dr. Miguel Beato

Department

Gene Regulation, Stem Cells, and Cancer Department of the Centre for Genomic Regulation (CRG)

upf. Universitat Pompeu Fabra Barcelona

Agraïments

I am very grateful for all of the help I have received during the PhD. First and foremost, I must acknowledge my supervisor, Miguel Beato. In my time in the lab, he has remained very accessible, even when he had more responsibilities as director of the institute. Unlike perhaps many supervisors, he gave me the opportunities to explore almost any idea I had. Not all of my ideas were the best, and he knew that, but he gave me the freedom to try and fail (and fail frequently) on my own. I must acknowledge Roderic Guigó. I perhaps did not take advantage of Roderic's extensive experience as best as I could have, but being a part of his lab in general helped me stay current with bioinformatic topics. In addition to that, he funded my final year of study, which has been the most fruitful. In addition to my supervisors, my thesis committee included Eduardo Eyras and Fátima Gebauer (who replaced Raul Méndez in 2011), who gave me very good advice even when discussing a topic outside of their specialty.

Beyond my formal supervision, I have to credit all of my labmates who have made possible my transformation from software engineer to capable experimentalist. This starts with Roni Wright. She taught me all the basics of culturing cells, and performing basic assays like Western blots or DNA gels, and also gene cloning techniques and transfection/infection of our cells. Jofre Font, Michael Wierer, Diana Reyes, Marija Kundakovic, Guille Vicent, Flopy Ogara, Priyanka Sharma, Alessandra Ciociola, Alejandro Lagreca, Silvina Nacht, and François Le Dily also helped me a lot in that respect. I learned extensive details about how to perform chromatin immunoprecipitations (ChIPs) from Cecilia Ballare. Other people that have benefitted my lab education through informal discussion (in no particular order) include João Tavanez, Bernhard Pätzold, Anna Corrionero, Camilla Ianonne, Gaetano Verde, Christos Gekas, Marina Garcia, Laura Gaveglia, Martin Lange, Payal Jain, Malte Beringer, Paula Pisano, Rory Johnson, Sophie Bonnal, Bill Keyes, Cristina Militti, Ramón Tejedor, Lluis Morey, Alessio Bava, Sergio Barberan, Luisa Vigevani, Luciano Di Croce, Juan Valcarcel, Karthik Arumugam, João Frade, Esteban Rozen, Adam Klosin, Alex Santanach, Elena Martin, Jordi

Hernández, Bruno Di Stefano, Laure Weill, Timo Zimmermann, Ben Lehner, Thomas Graf, Matteo Pecoraro, and Elias Bechara. Also I must thank people that contributed a wealth of information towards furthering my bioinformatic skills. This includes Giancarlo Castellano, Daniel Soronellas, João Curado, Christoforos Nikolaou, Debayan Datta, Hagen Tilgner, Colin Kingswood, Sonja Althammer, Julien Lagarde, Sarah Bonnin, Sarah Djebali, Emilio Palumbo, Brian Raney, Jim Kent, Anna Vlasova, Tobias Warnecke, David Haussler, Guillaume Filion, Alessandra Breschi, Barbara Uszczynska, Marco Mariotti, Dmitri Pervouchine, Raik Grünberg, Almer van der Sloot, Peter Vanhee, Gireesh Bogu, Panagiotis Papasaikas, Micha Sammeth, Cedrik Magis, Paolo Ribeca, Francisco Câmara, Ivan Junier, Davide Baù, Arnau Bria, Rodny Hernandez, Gabriel Gonzalez, Oscar Gonzalez, Judith Flo, Cedric Notredamme, Didac Santesmasses. I will also acknowledge Aitor Busquets, Isma De Mingo, Marc Gonzalez, Núria Janè, Romina Garrido, Imma Falero, Isabel Jurado, Diego Mellibovsky, Zoe Barbarà, Sharon Bel, Rafa Bayona, Carla Senozaín, Lucia Marucci, Lucia Russo, Giacoma Simonte, Francesco Sottile, Francesco Aulicino, Daniela Sanges, Pia Cosma, Livia Caizzi, Francesca Rapino, Mekayla Popoff, Valeria Di Giacomo, Valeria Giangarra, Alba Mas, Kadri Reis, Jia-Ming Chang, Anne Camapgna, Eric Verschueren, Maria Aurelia Ricci, Marc Friedländer, Tony Ferrar, Javier Delgado, Jae-Seong Yang, Solip Park, Mirko Francesconi, Krisztina Arató, Ilda Theka, Esther Lizano, Verónica Llorens, Emilia Szostak, Maria Lluch, Sophia Teichmann, Marcos Perez, Roxana Tovar, Chiara Di Vona, Luca Cozzuto, Jean-François Taly, Magalí Bartomeus, Eli Mateu, Elena Miñones, Carolina Gallo, Carolina Segura, Jo Aigner, Kiana Toufighi, Amy Curwin, James Cotterell, Alexandra Grippa, Luigi Aloia, Luigi Ombrato, Karl Wotton, Umberto Di Vicino, Eric Kallin, Julia Röwenstrunk, Marcus Buschbeck, Olivera Vujatovic, Sergi Repullo, Tian Tian, Valentina Schiavone, Silvina Catuara, Carla Bello, Tom Starke, Villy Michaki, Wassim Altarche, and Salvatore Cappadona in general for what I feel is a positive impact they each make on the CRG and why I have enjoyed working here.

Many of those that have helped me professionally, have also been good friends to me as well. The CRG and Barcelona can take credit for pro-

viding a wonderful environment, where it is easy to make friends and enjoy life. To that extent I will give special credit to my yoga instructor Alessa Bennaton and all of my spanish teachers in the past, most recently Alexandra Albuera. The city of Barcelona has given me back my love for running and I imagine I will find myself here again for more marathons.

My family is very important to me and they have supported all my life decisions, including moving to Europe and going back to school. This thesis is dedicated to the memory of my great-uncle Garland Petefish (1917-2008).

Thank you everyone.
Gracias a todos.
Gràcies a tothom.

Abstract

Genome-wide analysis of the nucleosome positioning and histone H1 isoform content of the T47D breast cancer cell line has found a number of observations, namely that with a gentle digestion of micrococcal nuclease (MNase), a nucleosome is visible just upstream of the transcription start site, in the region known as the “nucleosome-free region” (NFR). H1 isoforms bind to chromatin mainly in a redundant manner, but H1.2 and H1.3 show some specificity while H1.5 increases its binding dramatically after a progesterone stimulus. In the course of these studies, a general-purpose software package was developed for the manipulation and analysis of bigWig files, a data format for storing continuous signal data assigned to genome coordinates.

Resum

En el meu estudi genòmic sobre el posicionament de nucleosomes i sobre el contingut de les isoformes de la histona H1 en cèl·lules de càncer de mama T47D he dut a terme una sèrie d'observacions. Específicament he trobat que amb una digestió suau amb nucleasa micrococcal, es pot identificar un nucleosoma just abans del lloc d'inici de transcripció, en la regió coneguda com a "regió lliure de nucleosomes". També he vist que les diferents isoformes somàtiques de la histona H1 (H1.0-H1.5, H1x) s'uneixen a la cromatina de manera redundant, però que la H1.2 i la H1.3 presenten certa especificitat, mentre que la H1.5 mostra un augment de la unió generalitzat després d'estimular les cèl·lules amb progesterona. En el decurs de la meva recerca, he desenvolupat un programari general per la manipulació i l'anàlisi d'arxius amb format bigWig, un format per a l'emmagetzematge de dades de senyals continus al llarg de les coordenades del genoma.

Prefaci

This thesis touches on several different aspects of chromatin biology, in three chapters of results. After an introduction to chromatin biology and gene regulation and the kind of experiments that are done, I describe our work positioning nucleosomes in T47D breast cancer cells. This work is important because it draws attention to nucleosomes found in the badly named "nucleosome free regions" near gene transcription start sites. The second results chapter describes our work studying the genome-wide binding of linker histone H1 variants to chromatin. This study is notable because seven somatic variants of H1 exist in humans, and it is not understood why. Our study is also the first that analyzes the redistribution of H1 after a hormone stimulus to the cells.

A side effect of preparing these results was the creation of a utility for genomic continuous-valued signal data, whose function was originally split across a number of smaller more specific-use tools. This software: bwtool, has been released as open source to the bioinformatics community and its article recently published is attached as the third results chapter. Finally, I close with some additional discussion about how emerging trends and technologies may impact future directions of the field.

Sumari

Introducció

Gene regulation in eukaryotes

The simplicity behind the "central dogma" of molecular biology: that DNA is transcribed to RNA, which is then translated into protein, cannot explain everything. How can organisms with multiple types of cells: liver, muscle, skin, etc. have the same DNA in each of their cells? The answer is that the processes that dictate the conversion of DNA to eventual protein, known collectively as the regulation of gene expression, are highly controlled in order for the cell to produce the appropriate amount of protein. Particularly in eukaryotic cells, there are processes that regulate each step of a protein's production: before DNA is transcribed, regulation of transcription, regulation of the transcribed mRNA, regulation of the transport of the mRNA (including degradation of mRNA), regulation of translation, and finally post-translational modifications to the produced protein. Jacques Monod and François Jacob first described a system of gene regulation in *E. coli* where only in the presence of lactose are enzymes encoded by the lac operon produced [1]. The observation that specific proteins appear in response to specific stimuli was observed much earlier [2], but had no specified mechanism. Jacob and Monod's finding was the first to show precise modulation of the expression of lactose-metabolizing genes through factors acting as inducers and repressors. This discovery, as well as the observation in 1960 that the molting hormone ecdysone could induce formation of giant "puffs" in the chromosomes of diptera [3] paved the way for the study of steroid hormones and their ability to alter gene expression and chromosomal state. Induction and repression of genes are often controlled through cis-regulatory elements in promoters [4] or enhancers [5], both of which harbor specific sequences serving as substrates for the binding of transcription factor proteins. Promoters are generally located in the sequence immediately upstream of a gene, while enhancers

are located further, up to hundreds of kilobases away. Just prior to the discovery of these transcriptional regulators, it was first described that DNA is organized in conjunction with repeating units of histone proteins into a structure called chromatin [6]. It turns out that the organization of DNA into chromosomes and its more fundamental unit of chromatin is inextricably linked to the regulation of gene expression. Although it was seen as early as 1964 that histone post-translational modifications, specifically acetylation, could be linked to rates of transcription [7, 8], a mechanism for this regulation was not known. In the 1990s, the first genes encoding histone acetyltransferases (HATs), and histone deacetylases (HDACs) were cloned [9, 10], and the epigenetics field, which until then had been focused mainly on alternative splicing and transcription factors [11, 12], exploded. It was not long after this that ATP-dependent chromatin remodeling complexes, HAT/HDAC complexes, and transcription was all linked together in a concise way [13]. Although chromatin is a dynamic structure with a lot of unknown properties we have yet to uncover, the primary motivation when investigating chromatin is to extract meaningful connections between chromatin structural dynamics and gene regulation.

Chromatin structure

Chromatin has long been identified as the primary unit of DNA packaging in eukaryotic cells. To arrive at the DNA within the nucleus of a cell, one must unwrap the chromosome at different structural layers. Chromosomes are composed of a single molecule of double-stranded DNA, but this DNA is first wrapped and fastened around histone proteins to form nucleosomes [14] (Figure 1), like beads on a string. Furthermore, the string of nucleosomes is compacted in a semi-braided fashion to form structure known as the 30 nanometer fiber [15]. The 30 nm fiber twists and turns and ravels itself into larger super-structures such as chromatin loops and topologically associating domains (TADs) [16] eventually turning into the familiar structure of the chromosome, visible by light microscopy during mitosis and studied by cytogeneticists since the days of Walther Flemming [17] (see Figure 2).

The most rudimental purpose of chromatin is the compaction of

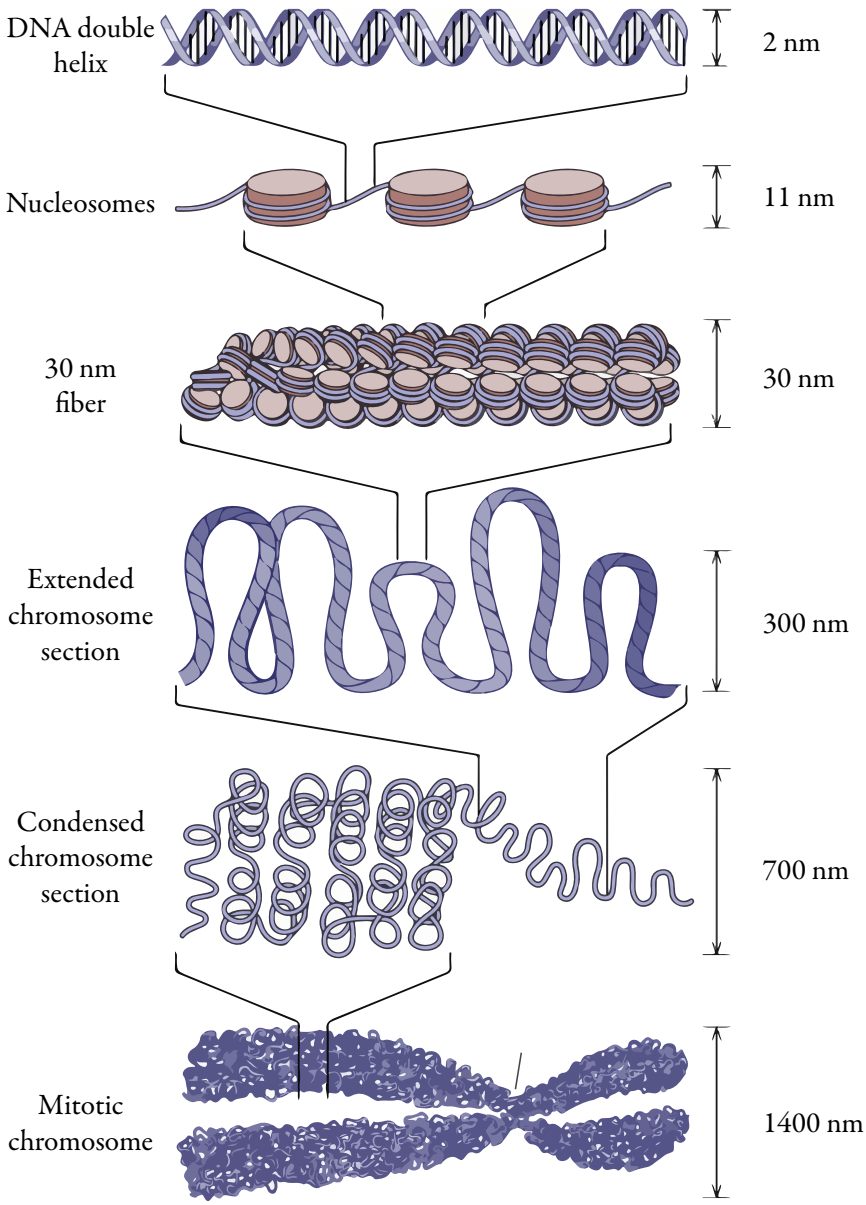


Figure 1: Chromatin at different magnifications. (From Jansen & Verstrepen, 2011 [205]).

DNA within the cell nucleus and its protection against mutation [18]. The wrapping of DNA around histone proteins and its further compaction reduces the length of the genome and its general accessibility of mutagens. The issue of accessibility does have additional consequences, namely, that in order

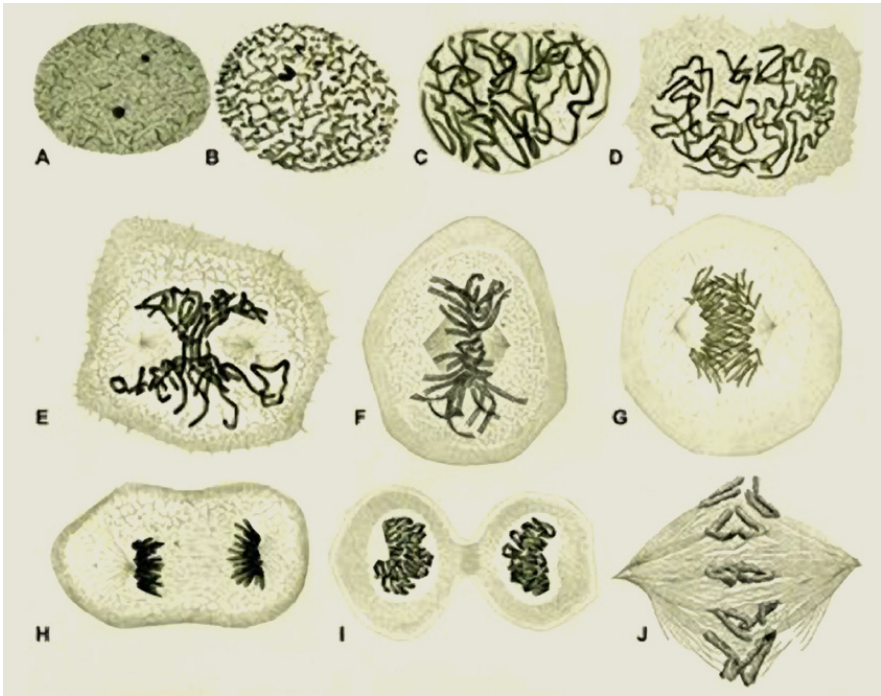


Figure 2: Chromosomes, as seen and hand-illustrated by Walther Flemming [206].

to read the information stored in chromatin, many molecular complexes are required to open and close the chromatin fiber, before, during, and after DNA transcription. In this context, post-translational modifications (PTMs) of the basic chromatin proteins introduce mechanisms for chromatin-binding proteins to act on or avoid specific regions of the DNA. The study of these specific non-genetic factors' roles in chromatin and their subsequent effect on phenotype form a large part of the field of epigenetics. The other part is accounted for by the modification of the DNA bases, particularly methylation of the Cytosines in CpGs that can be easily propagated through cell division.

Nucleosomes

The fundamental unit of chromatin is the nucleosome, and the fundamental unit of the nucleosome is the nucleosome core particle. Adding the surrounding variable-length linker DNA, as well as the linker histone H1 to the core particle results in the nucleosome. The core particle is composed of

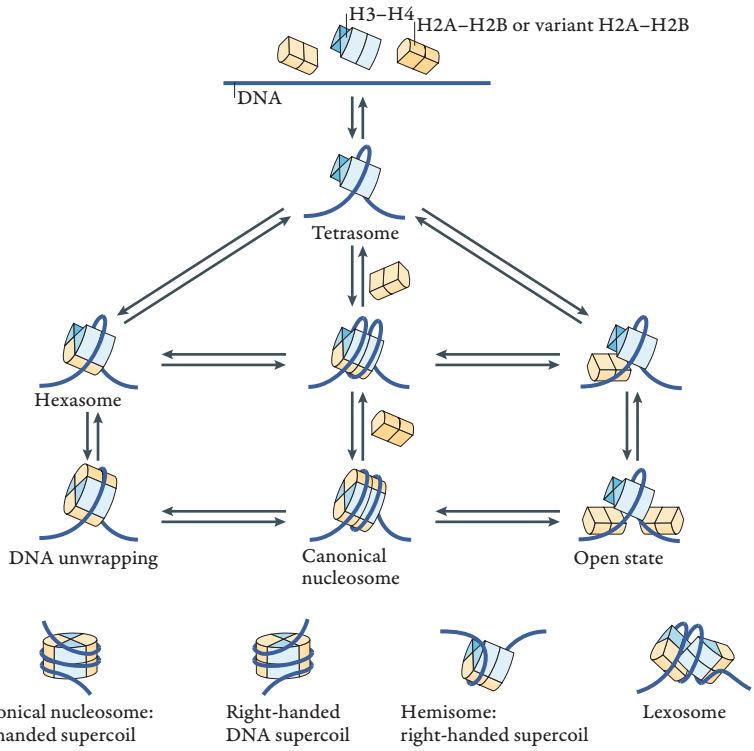


Figure 3: Canonical and alternate nucleosome structures. (from Luger, *et al.* 2012) [207].

147 base pairs of DNA wrapped around an octamer made of two each of the "core" histone proteins H2A, H2B, H3, and H4. With the full set of histones and the DNA in a left-handed coiled configuration, this is known as the "canonical nucleosome". It is generally what is thought of when one refers to a nucleosome, but other forms of nucleosomes exists (see Figure 3). Although the basic nucleosome core particle is a unit conserved in evolution as far back as yeast [19], the proteins involved are heavily post-transcriptionally modified and in some cases can be substituted by alternate variants of the protein entirely. Post-translational modifications (PTMs) of core histones, particularly in the tails of the core histones, have long been associated with differential gene expression [20, 21]. Numerous histone-modifying proteins exist that act to acetylate, methylate, phosphorylate, ubiquitinate, citrullinate proteins, and another set of proteins that remove those modifications. It is often necessary for a histone modification to exist in order to recruit the protein necessary for

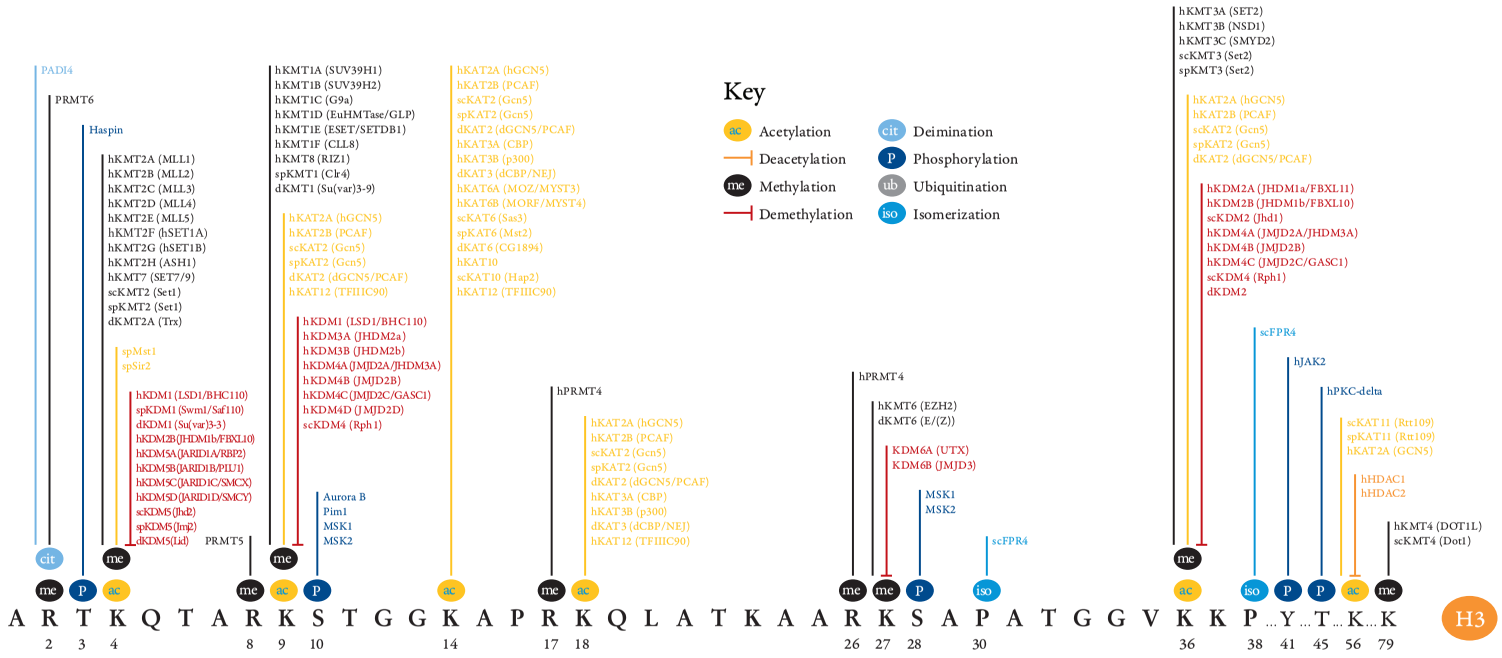


Figure 4: Histone H3 modifications (adapted from a figure on Abcam's website). There are a tremendous number of modifications to histone H3, even when making an incomplete list such as this one.

another histone modification. The temporal nature of these modifications form a series of events that then set the stage for transcriptional activation/repression and other processes. The extent that core histones are modified and the broad range of consequences these imply has led to the notion that the combinations of histone modifications can form what is known as the "histone code" [22, 23]. Histone H3 is perhaps the most modified histone. Figure 4 illustrates many of the known modifications of histone H3.

A more severe type of histone alteration exists when the entire protein is replaced by another altogether. Histone variants can carry out the same role or, as is often the case, replace the canonical histone at times, taking regulatory roles similar to the epigenetic roles of modified histones. Histone H4 is mostly invariant although in Drosophila there exists a replacement gene coding the same sequence [24]. Histone H2A is known for having several variants. In mammalian cells, macroH2A is enriched in the inactive X chromosomes resulting in increased recruitment of the Polycomb repressive complex 1 (PRC1) [25], while on autosomes has roles in the regulation of developmental genes [26]. Histone H3 is encoded by the H3.1, H3.2, and H3.3 genes, but H3.1 and H3.2 are considered both to be the canonical histone, while H3.3's deposition is associated with promoter and enhancer regulatory regions, particularly during gene activation [27]. When present, H2A.Z, like H3.3, disrupts the condensation of chromatin and is also associated with promoter regions. Additional specific roles for histone variants continue to be found, including the need for H2A.Z at nucleosomes in the vicinity of double-strand break repair sites (DSB sites) [28], or in the nucleosomes flanking active transcription start sites [29].

Histone H1

H1 is a eukaryotic protein, particularly important in the chromatin structure of metazoan species [30]. Despite yeast having the protein HH01, which has limited influence on chromatin structure and is mainly expressed during sporulation [31], histone H1 is mainly important in higher-order eukaryotes such as mammals. The effect of H1 on the folding of chromatin was first observed years ago [32] along with the observation that chromatin in

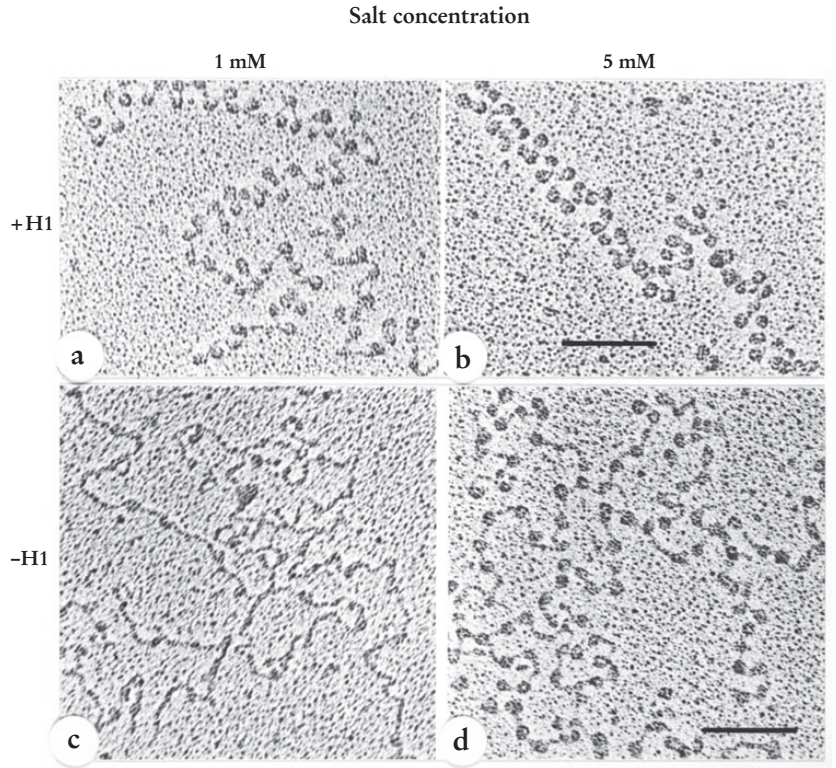


Figure 5: The effect of histone H1 on the structure of chromatin seen by Thoma, *et al.* (1979) [32]. With H1: (a) and (b), the structure is more ordered. Without H1: (c) and (d), the structure is much less ordered particularly at low ionic strength.

a solution with higher salt concentration will compact more. Histone H1 resides external to the nucleosome core particle, on the 10 bases entering and the 10 bases exiting the nucleosome. Compared with the core histones, the binding of H1 to the DNA is less stable [33], but it still is present in over 80% of nucleosomes[30]. For nucleosomes undergoing chromatin remodeling by ATP-dependent complexes such as SWI/SNF, ISWI, or Ino80, H1 is the first histone to be displaced [34]. Not all nucleosomes need histone H1 at all times, and it is frequently depleted in nucleosomes near transcription start sites. The lysine richness of histone H1 gives it cationic properties, and will increase the nucleosome repeat length until a 1:1 ratio of H1 and core histones is achieved [35]. Without H1, chromatin takes a linear form, resembling "beads on a string" [32] (Figure 5c, 5d). When H1 is present, even at

low ionic strength, chromatin will be a bit more ordered and have a zigzag appearance. With increasing ionic strength, more of the H1-binding regions will interact, eventually forming the solenoid superstructure (Figure 5b).

Metazoan H1 has a structure typically referred to as tripartite, meaning there is simply a globular domain flanked by an N-terminal domain (NTD) and a C-terminal domain (CTD). H1 is characterized by having many lysine residues, particularly in the CTD, and historically was named the "lysine-rich histone" [36]. H1 contains serine residues that are phosphorylated at low levels during the G1 phase of the cell, increasing through S and G2 until maximal levels of phosphorylation are reached in the late G2 phase [37]. Phosphorylation weakens H1's binding to chromatin, destabilizing the structure of chromatin locally. Other modifications of H1 have been seen. These include Poly(ADP-ribosyl)ation by PARP-1, and methylation of

```
H1.2            ----MSETAPAAPAAAPPAEKAPVKKKA-AKKAGGTP--RKASGPP-VSELITKAVAASK 52
H1.3            ----MSETAPLAPTIPAPAEKTPVKKK--AKKAGATAGKRKASGPP-VSELITKAVAASK 53
H1.4            ----MSETAPAAPAAPAEKTPVKKKA-RKSAGAAK--RKASGPP-VSELITKAVAASK 52
H1.5            ----MSETAPAETATPAPVEKSPAKKKATKKAAGAGAAKRKATGPP-VSELITKAVAASK 55
H1.1            ----MSETVPPAPAASAAPEKPLAGKKAKKPAKAAAASKKKPAGPS-VSELIVQAASSSK 55
H1.0            ----MTENSTSAPAAKP--KRAKASKKS-------------TDHPK-YSDMIVAAIQAEK 40
H1x             MSVELEEALPVTTAEGMAKKVTKAGGSAALSPSKKRKNSKKKNQPGKYSQLVVETIRRLG 60
                    : *  .  .:     : . .  .                 *   *:::. :

H1.2            ERSGVSLAALK-KALAAAGYDVEKNNSRIKLGLKSLVSKGTLVQTKGTGASGSFKLNKKA 111
H1.3            ERSGVSLAALK-KALAAAGYDVEKNNSRIKLGLKSLVSKGTLVQTKGTGASGSFKLNKKA 112
H1.4            ERSGVSLAALK-KALAAAGYDVEKNNSRIKLGLKSLVSKGTLVQTKGTGASGSFKLNKKA 111
H1.5            ERNGLSLAALK-KALAAGGYDVEKNNSRIKLGLKSLVSKGTLVQTKGTGASGSFKLNKKA 114
H1.1            ERGGVSLAALK-KALAAAGYDVEKNNSRIKLGIKSLVSKGTLVQTKGTGASGSFKLNKKA 114
H1.0            NRAGSSRQSIQ-KYIKSHYKVGENADSQIKLSIKRLVTTGVLKQTKGVGASGSFRLAKSD 99
H1x             ERNGSSLAKIYTEAKKVPWFDQQNGRTYLKYSIKALVQNDTLLQVKGTGANGSFKLNRKK 120
                :* * *   :  :         ::  : :* .:* ** ...* *.**.**.***:* :.

H1.2            ASGEAKPKVKKAGGTKPKKPVGAAKKPKKAAGGATPKKSAKKTPKKAKKPAAATVTKKVA 171
H1.3            ASGEGKPKAKKAGAAKPRKPAGAAKKPKKVAGAATPKKSIKKTPKKVKKPATAAGTKKVA 172
H1.4            ASGEAKPKAKKAGAAKAKKPAGAAKKPKKATGAATPKKSAKKTPKKAKKPAAAAGAKK-A 170
H1.5            ASGEAKPKAKKAGAAKAKKPAGAT--PKKAKKAAGAKKAVKKTPKKAKKP-AAAGVKKVA 171
H1.1            SSVETKPGASKV--ATKTKATGASKKLKKATGAS--KKSVK-TPKKAKKP---AATRKSS 166
H1.0            EPKKSVAFKKTKKEIKKVATPKKASKPKKAASKAPTKKPKATPVKKAKKK--LAATPKKA 157
H1x             LEGGGE---RRGAPAAATAPAPTAHKAKKAAPGAAGSRRADKKPARGQKP--EQRSHKKG 175
                              .   :   **.   :  .:       : :*       * .

H1.2            KSPKKAKVA-KPKKAAKS--AAKAVK-----PKAAKP-----KVVKPKKAAPKKK- 213
H1.3            KSAKKVKTP-QPKKAAKSPAKAKAPK-----PKAAKPKSGKPKVTKAKKAAPKKK- 221
H1.4            KSPKKAKAA-KPKKAPKSPAKAKAVK-----PKAAKPKTAKPKAAKPKKAAAKKK- 219
H1.5            KSPKKAKAAAKPKKATKSPAKPKAVKPKAAKPKAAKPKAAKPKAAKAKKAAAKKK- 226
H1.1            KNPKKPKTV-KPKKVAKSPAKAKAVK-----PKAAKARVTKPKTAKPKKAAPKKK- 215
H1.0            KKPKTVKAK-----------PVKASK-----PKKAKP--VKPKAKSSAKRAGKKK- 194
H1x             AGAKKDKGG-------------KAKK-----TAAAGGKKVKKAAKPSVPKVPKGRK 213
                  .*. *              ** *     .  *        .  .   . * :
```

Figure 6: Alignment of human somatic histone H1 variants. Highlighted in green is the globular domain, as annotated in each case by UniProt. Consensus residues are indicated by asterisks below the H1x sequence. The divergence of the H1.0 and H1x proteins distracts a bit from the nearly perfect conservation in the globular domain of the remaining isoforms.

Isoform	Official Gene ID	Mouse Homolog	Human Locus	Length (AA)	Chromatin condensation	Chromatin affinity
H1.0	H1F0	H1f0	chr22 (q13.1)	194	Medium	Medium
H1.1	HIST1H1A	Histh1a	chr6 (p22.2)	215	Low	Low
H1.2	HIST1H1C	Histh1c	chr6 (p22.2)	213	Negative	Medium
H1.3	HIST1H1D	Histh1d	chr6 (p22.2)	221	Medium	Medium
H1.4	HIST1H1E	Histh1e	chr6 (p22.2)	219	High	High
H1.5	HIST1H1B	Histh1b	chr6 (p22.1)	226	Medium	High
H1x	H1Fx	H1fx	chr3 (q21.3)	213	High	Low

Table 1: Human somatic histone H1 variants, and some general properties of each. Chromatin condensation was characterized using TMAFM (Tapping Mode Atomic Force Microscopy), using minichromosomes assembled in preblastodermic *Drosophilia* embryo extracts (DREX), with or without the presence of each H1 variant. Chromatin affinity was measured by the amount of each H1 required to affect the nucleosome spacing in DREX-assembled minichromosomes, seen on an MNase digestion ladder [50].

lysines by G9a/KMTC1 and Glp1/KMT1D [38], or citrullination [39].

Histone H1 Variants

The notion that multiple genes encode variations of histone H1 in vertebrates is not a new one. In 1966 the proteins later to be known as variants of H1 were fractionated from calf thymus tissue [40]. Humans have seven somatic (H1.0-H1.5, and H1x), and several germline-only variants. Unlike the variants of histone H3: H3.1, H3.2, and H3.3, which only vary in a few residues, histone H1 variants differ greatly in their amino-acid sequence, particularly in the CTD (Figure 6). Table 1 lists some general properties of the variants. Early studies in tobacco [41, 42], chicken DT40 cells [43], and mice [44], showed that knocking out one variant would raise the expression of the other variants, raising suspicions that the variants serve largely a redundant role. Later, it was found that knocking out individual mouse H1 isoforms had little effect, while a triple knockout of H1.2/H1.3/H1.4 was lethal in embryos by mid-gestation [44, 45].

H1.0

Among the somatic variants, H1.0 is the most divergent in its amino acid sequence. This is due to it being part of a more ancient lineage of H1 histones that diverged from the main group of H1 proteins before vertebrates. H1.0 descends from the same group as H5 [46], a H1-like protein present in mainly transcriptionally-inactive chicken erythrocyte cells. H1.0 accumulates in terminally-differentiated cells, and within these cells is associated with inactivated genes [47]. It's for this reason that H1.0 has been often thought of as having an intermediate form between that of H5 and other H1 histones [48].

H1.1

H1.1 is only expressed in certain tissues [49], most notably testis, thymus, spleen, lymphocytes. In vitro, it is one of the weakest chromatin condensers [50], but has a medium affinity to chromatin [51]. In IMR90 cells, genome-wide DamID binding assays showed that H1.1 had the most specific binding pattern compared to H1.2-H1.5, which by contrast displayed more redundancy [52].

H1.2

Along with H1.4, H1.2 is possibly expressed in all human cells, raising the possibility it serves a very important function [53]. It has a short C-terminal domain, which may contribute to its short binding periods. H1.2 is a histone with proposed function outside of the nucleus. H1.2 has been seen to localize to the cytoplasm in response to X-ray induced DNA damage, and therapeutic treatment of leukemia, and could be involved in the regulation of apoptosis through Bak-mediated mitochondrial release of cytochrome C [54].

H1.3

Histone H1.3 has been seen in immunostains to associate with euchromatin [55]. H1.3 is also known to be expressed at very low levels, and is depleted in actively-transcribed chromatin [56].

H1.4

Histone H1.4 is associated with heterochromatin and has strong condensing

properties. Along with H1.2, it is seen in every human cell tested [53]. Evidence exists that H1.4 is involved in the cell cycle and cell death [57].

H1.5

H1.5 has a long C-terminal domain, contributing to longer residence times on chromatin versus variants H1.0-H1.3. There is somewhat conflicting evidence over H1.5's role in stem cells. On the one hand, H1.5 was seen to have enriched binding in membrane or membrane-related proteins in terminally-differentiated cells (IMR90 fibroblasts) versus little or no binding in any major gene family in undifferentiated cells (H1 hESCs) [58]. On the other hand, pluripotent cells (embryonic stem cells and induced pluripotent keratinocytes) had increased levels of H1.5 compared with differentiated cells [59]. In breast and colorectal cancers, H1 proteins are expressed differently than in normal tissue, but H1.5 was shown to be mutated [60].

H1x

H1x is the most divergent H1 subtype in terms of its sequence. It has been shown to have cell cycle-dependent distribution in the nucleus: accumulating in the nucleoli in the G1 phase, but evenly distributed in S and G1 phases [61]. H1x has also been seen to have higher expression versus H1.0 in neuroendocrinal tumors compared to normal tissue [62].

Intermediate chromatin structure

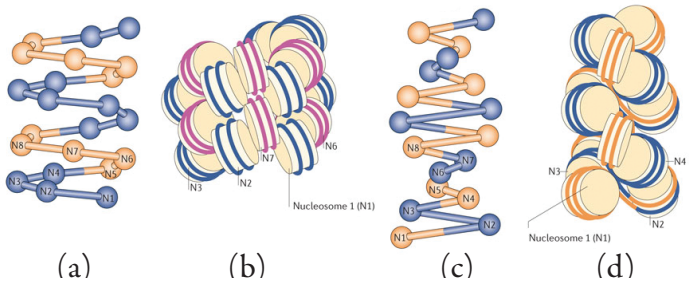


Figure 7: The two main proposed conformations of 30 nm chromatin fiber. In the solenoid model: (a) and (b), the histones of adjacent nucleosomes interact. In the zigzag model: (c) and (d), histones from alternating nucleosomes more commonly interact (from Luger, *et al.* 2012)[207].

The intermediate chromatin structure generally refers to the structure seen in microscopes between the 10 nm "beads-on-a-string" nucleosomes and whole chromosomes. As chromatin condenses, it forms a "30 nm fiber" structure. The exact conformation of this structure is still the subject of controversy. Some believe the "solenoid" structure proposed in the late 1970s by Finch and Klug to be correct, while others believe zig-zag model is more accurate. Figure 7 (from [63]) shows different proposed 30 nm structures. The trouble could lie in varying experimental conditions: potential artifacts introduced by cross-linking and fixation for microscopy, or the levels of magnesium and chloride [63]. The 30 nm fiber is a difficult structure to elucidate, which is why even very recently, there remain studies that doubt the existence of a 30 nm structure at all [64, 65]. Our view on the subject is relatively open-minded, however for the purposes of this thesis we assume its existence in vivo. Recently, in vitro work on the 30 nm fiber using Cryo-electron microscopy has shown a zigzag conformation in the 30 nm fiber at two different nucleosome repeat lengths: 177 bp and 187 bp [66]. Both show a strong tendency for the stacking of nucleosomes to form left-handed double-helical twists. Although the work was again done in vitro, it is the first to sufficiently capitalize on the

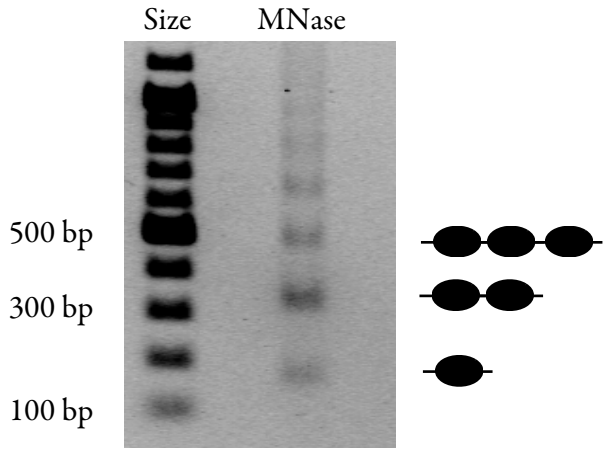


Figure 8: An example MNase digestion ladder, in this case done with chromatin reconstituted using recombinant, purified core histones and plasmid DNA, and facilitated by incubating with assembly factors (McNAP and the ACF complex). As is often the case with a light digestion of MNase, a heavier band is seen for the dinucleosome molecular weight than the mononucleosome.

Cryo-EM technique. Cryo-EM, in theory, has the potential to answer the 30 nm question once and for all, because unlike X-ray crystallography, it allows the observation of samples without staining or fixing, preserving their native physiological environment.

Nucleosome Positioning

There are several motivating factors when considering the reasons it is desirable to obtain positions of nucleosomes, and not just accept that they are present at arbitrary points on the DNA. First, positioning of nucleosomes is a major determinant in the precision of gene regulation [67, 68]. Second, the positions of nucleosomes are highly influenced by the underlying sequence of DNA it occupies. These sequence preferences are important, but not the only factor in positioning. The model of "statistical positioning" proposed by Roger Kornberg, hypothesizes that nucleosome positions are influenced by neighboring nucleosomes, and that perhaps only a subset of nucleosomes are well-positioned [69]. To gather nucleosome positions experimentally, assays exist to isolate DNA bound to by nucleosomes. The use of micrococcal nuclease (MNase) to cleave chromatin at DNA linkers has been used to isolate nucleosome-protected DNA since the early years of chromatin biology. It was first shown that an endogenous nuclease in rat liver cells could cleave DNA into uniform sizes, and multiples of that size [70]. Not long after this, MNase was found to cleave DNA in regular 200 bp intervals [71], forming a ladder of evenly-spaced bands seen by electrophoresis (Figure 8). The exonucleic activity of the MNase enzyme cleaves preferentially in nucleosome linkers, but sustained incubation would cause the enzyme to then endonucleically digest DNA, first resulting in 166 bp fragments, and eventually 146 bp fragments, the former due to the presence of histone H1 [72, 73].

A long-standing concern of using the MNase enzyme is that the enzyme itself displays biases in its cleavage site [74], which is typically centered at an AT dinucleotide. Additional ways of cleaving nucleosomes have been found, including using chemicals such as Methidiumpropyl-EDTA-iron(II) ($C_{34}H_{39}N_7O_8$) [75, 76], cuprous phenanthroline ($C_{12}H_8CuN_2$) [77], but these are more sensitive and difficult to use in vivo, and anyway have biases of their

own. To limit the effect of the bias, controls in MNase experiments include the comparison of MNase-digested nucleosomal fragments to MNase-digested free DNA [78], simulated MNase-digested chromatin based on the MNase cleavage consensus and corresponding sites in the genome [79], or MNase-digested chromatin followed by chromatin immunoprecipitation using an antibody against a core histone [80]. Despite the concerns posed by the enzyme, it is still generally accepted to be the best method to position nucleosomes [81].

Prior to MNase-sequencing the classical protocols for nucleosome positioning include indirect end-labeling [82], primer extension [83], hydroxyl radical footprinting [84], or monomer extension [85]. These methods are limited to positioning one, or in some case just a few, nucleosomes. This has resulted in several nucleosome models including the MMTV promoter [86] or the *H. polymorpha* MOX promoter [87].

Translational vs. rotational positioning

As a matter of nomenclature, it is important to distinguish the two main types of nucleosome positioning. Quite simply, translational positioning is defined by the locus of DNA occupied by a histone octamer in the genome. An MNase-sequencing experiment provides this information. Rotational positioning defines the conformation of the histone octamer within the DNA that wraps it. A nucleosome may have multiple rotational positions without moving translationally. Rotational positioning is harder to determine than translational positioning, but with careful DNase I digestion [88], it is possible to see 10 bp repetitions in a high-resolution gel at a single nucleosome locus, corresponding to different rotational phases of the nucleosome. It has also been possible to see rotational phasing with a combination of MNase-seq and DNase-seq, albeit by averaging signals from loci genome-wide [79].

Nucleosome positioning *in vitro*

Apart from the in vivo work on positioning nucleosomes to specific loci, studies have been done using artificial DNA sequences to analyze nucleosome affinity and its capacity to position and reposition. Some of this work

overlaps with the research mentioned in the histone H1 section, but there has also been work done specifically for nucleosome positioning. The "601 sequence" is the primary example of a DNA template engineered specifically for high-affinity, sequence-directed positioning of nucleosomes [89], constructed using a SELEX-based method of selecting preferred nucleic acid ligands to a target protein or RNA from a pool of random oligonucleotides [90]. The "601 sequence" has been used in many studies to obtain precise kinetic data concerning various chromatin remodelers like ACF [91], or histone chaperones like NAP1 [92]. It has even been crystalized with bound nucleosomes [93].

Predicting nucleosome positions

Very related to the fine-scale studies of in vitro nucleosome positioning, is the computational field of nucleosome position prediction. Predicting nucleosome positions from underlying DNA sequence has been a long-term goal of the field, because it would represent as much as possible a true understanding of why nucleosomes are bound to certain sequences as not others. A slightly overzealous attempt at providing a "genomic code" for nucleosome positioning was proposed in 2006 [94], which provided a computational algorithm for determining nucleosome positions based on a probabilistic model, the heart of which was a probability PN(S) for each 147-base sequence a predefined background probability distribution PB(S) with Score(S) = log(PN/PB). This algorithm found that 54% of the predicted stable nucleosome positions were within 35 bp of experimental positions. The article garnered an impressive amount of attention [95], but left a lot of room for improvement. An updated version of the algorithm was released and changed the model by introducing a different background probability: PL(S), based on 5-mer sequences found in nucleosome linkers, thereby allowing the model to capture both nucleosome and non-nucleosome favored positions [96]. Other software has been created to predict nucleosome positioning directly from DNA sequence using preconfigured mathematical models: NuPop [97], FineStr [98], or the method by van der Heijden, *et al.* (2012) [99] are all examples. SymCurv [100], also directly uses DNA sequence, but uses DNA bendability

and curvature with the additional consideration that because the DNA helix wraps itself around the histone octamer 2.5 times, that the points in the helices lying next to each other will be subjected to the same physical forces. Perhaps not all nucleosomes, but particularly well-positioned ones, may benefit from an inherent symmetry in the bendability in the DNA at these positions.

Genome-wide positioning *in vivo*

The statement that the primary determinants of nucleosome positioning were the intrinsic control sequences in the DNA provoked a strong reaction from the nucleosome community [101-103]. As important as the sequence of underlying DNA is to the positioning of nucleosomes, extrinsic factors remain highly influential. The obvious suspects at work in this case are the various DNA/chromatin binding proteins and complexes. One example, the zinc finger protein CTCF, involved in a multitude of regulatory processes like transcriptional activation/repression [104], insulation[105], chromatin looping [106], and genetic imprinting [107], has also been shown to influence nucleosome positioning [108-110]. Many other associations to experimentally obtained nucleosome positions have been made. Nucleosomes are

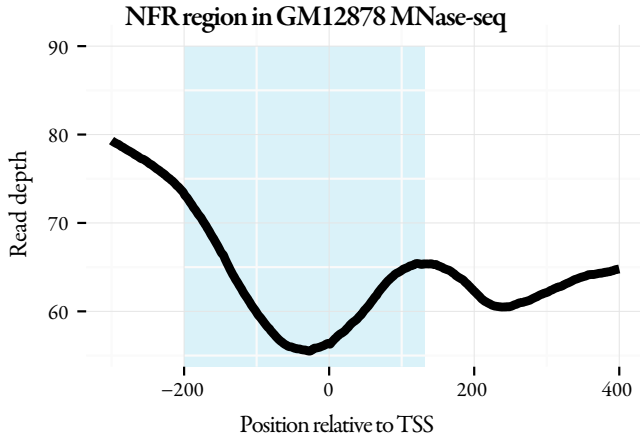


Figure 9: A typical "nucleosome free" region (NFR) seen in ENCODE GM12878 fibroblast MNase-sequencing data [209] (highlighted in blue), derived from aggregation at the TSS for the 20,330 protein-coding genes annotated in GENCODE v17. The +1 nucleosome is clear, preceded upstream by a marked depletion of nucleosome occupancy of around 300 bp, and including the TSS itself.

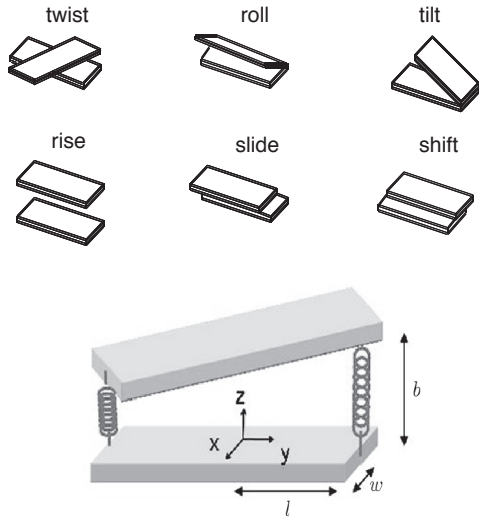


Figure 10: Different ways of deforming a basepair dinucleotide stack (from Ghorbani & Mohammad-Rafiee, 2011) [210].

well positioned at exon/intron boundaries [111-113], nucleosomal DNA is preferentially methylated compared to flanking DNA [114], and nucleosomal DNA harbors fewer single nucleotide polymorphisms (SNPs) than non-nucleosomal DNA [18].

Other results from mammalian genome-wide nucleosome positioning experiments indicate different cell types exhibit widely different spacing of nucleosomes [115], that around 8-9% of nucleosomes are strongly-positioned, and tandem repeat sequences naturally occur in the human genome (e.g. on chromosome 12) that cause very well-positioned arrays of nucleosomes [79]. The first dynamic system of nucleosome positioning to be reported using mammalian cells was done with mouse embryonic stem cells and neural and embryonic fibroblast progenitors [110]. They found a 5-7 bp increase in nucleosome repeat length (varying locally), as well as correlations between nucleosome occupancy and specific histone methylations and acetylations.

The "nucleosome-free" region

The most defining feature of a nucleosome positioning experiment is not actually a nucleosome position at all, rather the lack of nucleosome occupancy in the region of roughly 250 bp comprising the 200 bp upstream of a

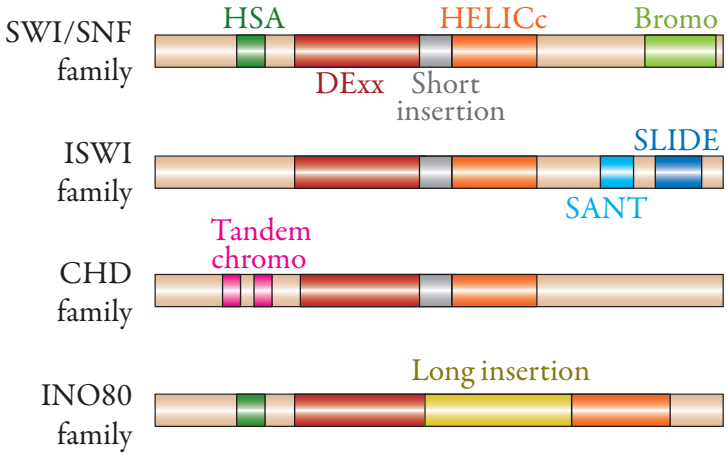


Figure 11: Structure of major ATP-dependent chromatin remodeling families and the protein domains they are characterized by (from Clapier & Cairns, 2009) [211].

transcription start site, and 50 bp downstream. This region has come to be known as the "nucleosome-free region" (NFR) [116-121], though sometimes it is also referred to as the "nucleosome-depleted region" (NDR). The NFR is seen in virtually every genome-wide nucleosome positioning experiment to date. Figure 9 shows a prototypical NFR region using nucleosome-positioning experiments of human blood cells.

The DNA sequence in NFR regions has interesting characteristics. NFR regions are sensitive to other nucleases, such as DNaseI, and are rich in transcription factor binding sites, of which many have known sequence motifs. The MNase enzyme used in nucleosome positioning experiments is biased to cleave at AT dinucleotides, which the NFR is enriched in, along with AA dinucleotides. The sequence attibutes of the NFR have an impact on the physical structure of the promoter. The amount of rolling, tilting, sliding, shifting etc. between bases stacked together in dinucleotides or trinucleotides is known from thermodynamic properties of nucleic acids. These properties can be extrapolated to longer sequences with biophysical modeling, to the extent that the unique DNA bendability and deformability found in promoter regions has been used to successfully predict novel promoter regions [122].

Chromatin dynamics

PTM	Domain	Protein	Functions
H3K4me0	PHD	BHC80	LSD1 complex
		AIRE	Autoimmune regulation
	WD40	WDR5/WDR9	HAT
	ADD	Dmnt3L	DNA methylation
H3K4me	Chromo	CHD1	ATPase, chromatin remodeling
	PHD	RAG2	Recombination
		ING2	HDAC
		BPTF	ATPase
		TAF3	TFIID
		PHF2	H3K9 demethylase
		ING4	
		YNG1	
		PHF8	
	Tudor	JMJD2A	
		JMJD2C	
		Sgf29	
H3K9me	Chromo	HP1	Heterochromatic mark
		CDY, CDYL, CDYL2	Repressor of REST
	PHD	SMCX	Demethylation
	Tudor	TDRD7	
		UHRF1	
	WD40	EED	PRC2 activity
		LRWD1	DNA replication
	Ankyrin repeats	G9a/GLP	Methyltransferase
H3K27me	WD40	EED	Polycomb repression
		LRWD1	DNA replication
	Chromo	PC	PRC1 complex
		CDY, CDYL, CDYL2	
		CBX7	Polycomb repression
		MPP8	
H3K36me	Chromo	Eaf3	HDAC
		MSL3	Dosage compensation
		MRG15	Splicing

PTM	Domain	Protein	Functions
H3K36me	PWWP	DNMT3A	DNA methylation
		BRPF1	Histone acetylation
		NSD1, NSD2, NSD3	Histone methylation
		MSH-6	DNA mismatch recognition
		N-PAC	Transcription elongation
H1K26me	MBT	L(3)MBTL1	Chromatin lock
	WD40	EED	Inhibits PRC2 methyltransferase
H3R17me	Tudor	TDRD3	Transcription activation
H3S10ph	(Gcn5)	Gcn5	Histone acetylation
H3Y41ph			Exclude HP1α binding
H2bK120ub/ H2BK123ub		Cps35	H3K4 methylation
H3K14ac	Tandem PHD	DPF3b	Chromatin remodeling
	Tandem Bromo	Rsc4	Chromatin remodeling
	Bromo 2	Polybromo	Chromatin remodeling

Table 2: Examples of histone modification readers and their recognizing domain, and the reader's function. (Adapted from Yun *et. al* 2011) [212].

ATP-dependent Chromatin Remodeling

Chromatin remodeling is categorized into two groups: the first, which requires the hydrolysis of ATP, and the second, which does not. The latter involves the enzymes mentioned previously, that catalyze covalent modifications to histones. These changes often result in changes in the conformation of the chromatin, particularly when in concert with other modifications or on histone variants. The other type of chromatin remodeling involves the sliding, twisting, or looping of DNA around nucleosomes, and eviction and exchange of histones from the DNA. There are five main families of protein complexes that make up this group: SWI/SNF, ISWI, CHD, INO80, and SWR1. Each contains an ATPase domain, but they have different protein-binding domains (Figure 10), each serving very different purposes [123]. Of the five families, the SWI/SNF and ISWI are the families that have been studied the most. Other than their different biological roles, they can also be characterized in their contrasting mechanisms of sliding the nucleosome.

Briefly, many ISWI remodelers regularly space, or transitionally phase, nucleosomes from an initially random set of nucleosome positions, while SWI/SNF remodelers tend to disrupt the order of a positioned array of nucleosomes [124]. Related to the SWI/SNF complex in yeast is the RSC complex [125], which interestingly forms a complex with a partially-unwound nucleosome, and can evict the histone octamer of neighboring nucleosomes [126] and recruit the Gal4 transcriptional activator [127].

Histone modifying enzymes and readers

Post-translational modifications to histones occur through the mechanisms of various enzymes. Already mentioned were the HAT/HDAC enzymes involved in histone acetylation/deacetylation. Many of these modifications allow the binding of specific proteins, also known as "readers", the result of which leads to various consequences. Table 2 lists examples of modifying enzymes, their associated modification, and examples of associated proteins that make use of a histone modification.

Steroid Hormone Receptors

Steroid hormone receptors are one of three classes of nuclear receptors: transcription factors activated by specific ligands, regulating the expression of target genes [128]. The second class is the thyroid/retinoid family of receptors: thyroid receptor (TR), vitamin D receptor (VDR), retinoic acid receptor (RAR), and peroxisome proliferator-activated receptor (PPAR). The third class of nuclear receptors are known as the orphan receptor family, and have unknown ligands.

Steroid receptors bind to DNA as homodimers, and their ligands, which they bind to with high affinity, are from endogenous endocrine sources. Members of this group include the glucocorticoid receptor (GR), mineralocorticoid receptor (MR), estrogen receptor alpha and beta, androgen receptor (AR), and the progesterone (PR) receptors. Although these receptors can have major roles in puberty and development [129], much of the focus of their study is on their effects of cancer cell proliferation [130-132]. Com-

mon to steroid hormones are the two main domains of the protein: a central DNA-binding domain (DBD), and a C-terminal ligand-binding domain (LBD). Also common are several nuclear localization signals in the C-terminus, hinge domains, and the LBD [133]. For AR, GR, MR, and PR, the same sequence of DNA is recognized by the receptor for binding. These hormone responsive elements (HREs) are usually composed of two palindromic hexanucleotides, each of which can bind a receptor monomer, and are separated by three non-conserved nucleotides.

The Progesterone/MMTV Model for Studying Chromatin

Before the availability of genome-wide methods like RNA-seq, microarrays, or ChIP-seq, the MMTV promoter (mouse mammalian tumor) was used extensively for the analysis of the binding of steroid hormone receptors and subsequent activation of the promoter [134]. Originally known as the "Bittner virus", John Bittner proposed in 1936 that cancers were being passed from mother to progeny mice through milk [135]. Years later, isolating the MMTV gene in plasmids [136] led to the discovery that its activation could be mediated by hormone induction [137]. Soon, HRE sequences in the MMTV promoter were described for the binding of glucocorticoid and progesterone receptors [138-140]. Then, nucleosomes were positioned to the MMTV promoter [68]. The MMTV promoter has been infected into a number of different cell models including cell lines, Saccharomyces cerevisiae [141] and made the basis of minichromosomes added to Drosophila melanogaster cell extracts [142].

The T47D cell line is a mammary ductal tumor cell line, notable for having an active progesterone receptor [143]. The combination of T47D cells and the MMTV promoter as T47D-MTVL cells has provided a useful model for the study of chromatin in the context of the progesterone. The progesterone receptor (PR) provides a mechanism for studying chromatin remodeling and gene regulation in two main ways. The first is called the "genomic pathway" because PR, ligated to progesterone or estrogen or in some cases non-ligated, binds directly to chromatin acting as a transcription factor and recruits

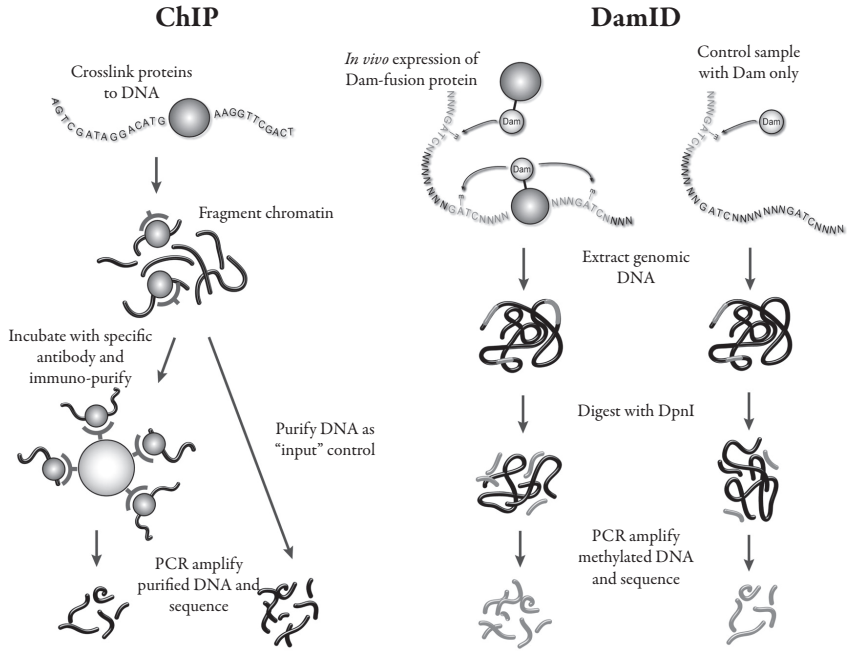


Figure 12: Schematic showing the ChIP and DamID methods and their negative controls (adapted from van Steensel *et. al*, 2005)[212].

chromatin-remodeling complexes. The second mechanism, called the "non genomic" pathway, concerns PR in the cytoplasm activating one or several signaling pathways including Src/p21/Erk [144] and Jak/STAT [145]. Combining both mechanisms results in roughly 2,000 genes being up-regulated by progesterone and 2,000 down-regulated genes [146]. ChIP experiments using T47D-MTVL cells have revealed many insights into the mechanisms of nucleosome remodeling during MMTV promoter activation.

Methods to study chromatin

Chromatin immunoprecipitation

By far the most widely used contemporary method to study chromatin is to use chromatin immunoprecipitation, also known as "ChIP" [147]. The ChIP

method essentially isolates DNA bound to a protein of interest (see Figure 11a). The method relies on an antibody with good specificity to immunoprecipitate the antigen target protein out of solution, along with a short fragment of DNA it had been bound to. To probe specific proteins, antibodies are raised against full-length targets or peptides unique to that target. Crosslinking chromatin with formaldehyde covalently fixes chromatin proteins and complexes to the DNA. In this way, proteins indirectly bound to chromatin are also precipitated if they are part of a complex that is bound. With the precipitated protein/DNA in hand, crosslinks are reversed (if they were used), and the DNA purified. Once the DNA is obtained, it can be used in PCR experiments along with designed PCR primer pairs. DNA microarrays can be used to hybridize the ChIP DNA in ChIP-chip [148], or more common in recent years, the DNA can be sequenced using high-throughput sequencing: the ChIP-seq [29, 149, 150]. ChIP-seq has been a fantastically popular method and has been used by consortia such as ENCODE to scour the genome using hundreds of antibodies against chromatin proteins or specific modifications of those proteins [151]. The ChIP-seq method is not with its drawbacks, however. The experiment depends heavily on the quality of the antibody, and comparing two ChIP-seq experiments using different antibodies is not trivial, l*et alo*ne hundreds of ChIP-seqs. One way to circumvent this problem is to fuse epitope tags to proteins of interest using tags such as FLAG [152], HA [153], or MYC [154]. These tags allow a single antibody to be used across multiple experiments targeting different proteins, but introduce problems associated with overexpressing those proteins. Nevertheless, the explosion of this data in recent years has had a profound sharpening of our view of chromatin: where and when various proteins are bound, and when certain proteins are modified.

There are several alternatives to ChIP-seq that offer ways of isolating different chromatin-bound material, or they are different methods achieving the same purpose. One method, DamID (DNA adenine methyltransferase identification) [155, 156], works by fusing the target protein with a DNA methyltransferase (see Figure 11b). When the target fusion protein binds to chromatin, the methyltransferase will methylate adenines in nearby GATC

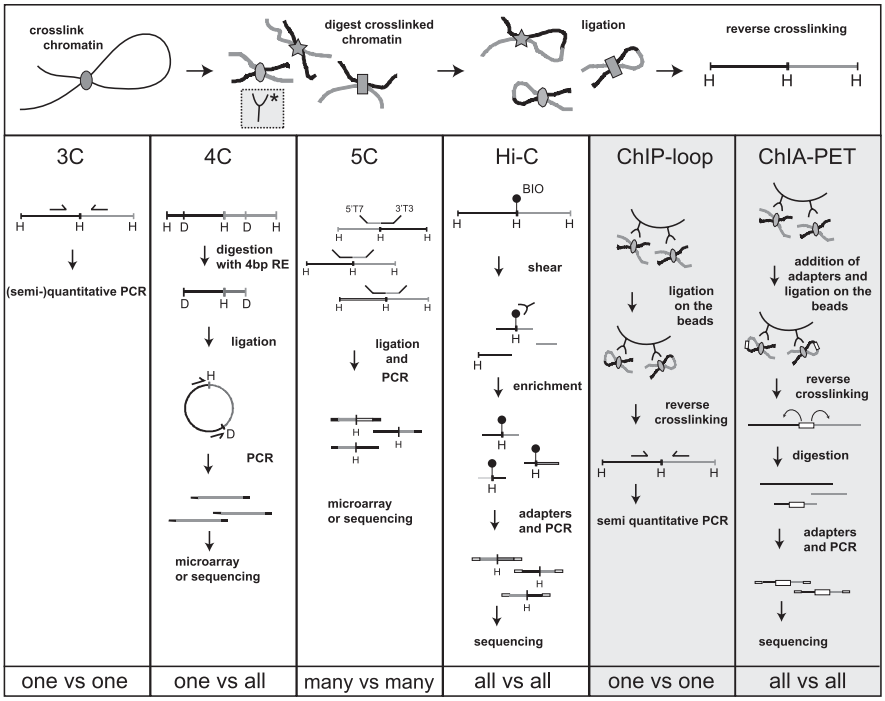


Figure 13: An overview schematic of various 3C (chromatin conformation capture) methods, from de Wit, *et al.* (2012) [175]. Apart from ChIP-loop and ChIA-PET, the methods begin by crosslinking chromatin, digesting with a specific restriction enzyme, and ligation in dilute conditions. 3C is for a specific locus, requiring PCR primers to target sequence near restriction sites to capture ligation junctions. With 4C (circular 3C) [214, 215], a second round of digestion and ligation is done to create circular DNA. With primers for the locus of interest and inverse PCR, the contacting sequences are found by sequencing. 5C uses a library of oligonucleotides containing the same restriction sites, and are hybridized to the 3C template. Pairs of oligos representing contacting fragments can be ligated together. Hi-C doesn't use a specific library of any kind, rather it relies on a the restricted fragments to be filled-in with biotin-labeled nucleotides, which can then be specifically pulled-down with streptavidin beads after ligation. This enriches the set of fragments containing restriction sites, and with paired-end sequencing it can be determined if the ligation involved interacting fragments. ChIP-loop [216] and ChIA-PET [217] are modifications chromatin immunoprecipitation protocols, allowing long-range interactions to be captured with the addition of a ligation step just after the bead enrichment step. Like ChIP, ChIP-loop and ChIA-PET use an antibody against a chromatin protein of interest and are therefore ways of capturing long-range DNA interactions to specific protein binding sites.

sequences. Because adenosine methylation does not occur naturally in eukaryotes, these methylations are exclusively attributed to the DamID method. These methylations can then be mapped using methylation-specific restric-

tion enzymes or antibodies. Methods that could be considered the "RNA cousins" of ChIP-seq have been established to find RNA or RNA-binding proteins that interact with chromatin. These include: CLIP [157], RIP [158-161], or ChIRP [162].

Nuclease sensitivity

While protein-based methods like ChIP-seq are certainly the most widely used tools in chromatin epigenetics, conceptually more simple are methods using nucleases to cleave chromatin DNA unprotected by histones or other bound proteins. Much of what we know about chromatin in the first 20 years of the field comes from the use of these nucleases. One of these nucleases: micrococcal nuclease, was discussed in the section on nucleosome positioning (see section 1.5). Another nuclease, deoxyribonuclease I (DNase I), is a human protein expressed in apoptotic cells that digests DNA. As a somewhat bulky enzyme, it cleaves DNA first in the most accessible regions of chromatin. A DNase I hypersensitivity assay will find regions of chromatin, called "hypersensitivity sites" that are not condensed like most chromatin [163]. These open regions of chromatin were soon associated with active genes [164] or enhancer regions [165]where transcription factors or polymerases need access to the DNA. Combined with high-throughput sequencing, DNAse-seq [166-168], has become another technique adopted by consortia like ENCODE for genome-wide characterization of open chromatin [169]. Another method, called FAIRE-seq (Formaldehyde-Assisted Isolation of Regulatory Elements) [170], has also been used for the same purpose as DNase-seq, and in some cases the two methods have been used together to obtain a stronger result [171].

Long-range chromosomal interactions

In the past decade several methods have emerged to find where regions of a chromosome is in contact with itself or another chromosome. This is interesting for two reasons: (1) the extent of the complexity between the 30 nm fiber structure of chromatin and the whole chromosome is not well character-

(a)

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

(b)

```
>gi|46452254|gb|AY585334.1| Sus scrofa cystic fibrosis transmembrane conductance regulator
(CFTR) mRNA, complete cds
Length=4449

 Score = 5453 bits (2751),  Expect = 0.0
 Identities = 4036/4449 (90%), Gaps = 6/4449 (0%)
 Strand=Plus/Plus

CDS: Putative 1       1     M  Q  R  S  P  L  E  K  A  S  V  V  S  K  L  F  F  S  W  T
Query                 133   ATGCAGAGGTCGCCTCTGGAAAAGGCCAGCGTTGTCTCCAAACnnnnnnnCAGCTGGACC  192
                            ||||||||||||||||||||||||||||||| |  ||||||||||||||||||||||||
Sbjct                 1     ATGCAGAGGTCGCCTCTGGAAAAGGCCAGCATCTTCTCCAAACTTTTTTTCAGCTGGACC  60
CDS:cystic fibrosis   1      M  Q  R  S  P  L  E  K  A  S  I  F  S  K  L  F  F  S  W  T

CDS: Putative 1       21    R  P  I  L  R  K  G  Y  R  Q  R  L  E  L  S  D  I  Y  Q  I
Query                 193   AGACCAATTTTGAGGAAAGGATACAGACAGCGCCTGGAATTGTCAGACATATACCAAATC  252
                            ||||||||||||| |||||||| ||||||||||||||||||||||||||||||||||| |||
Sbjct                 61    AGACCAATTTTGAGAAAAGGATATAGACAGCGCCTGGAATTGTCAGACATATACCATATC  120
CDS:cystic fibrosis   21     R  P  I  L  R  K  G  Y  R  Q  R  L  E  L  S  D  I  Y  H  I

CDS: Putative 1       41    P  S  V  D  S  A  D  N  L  S  E  K  L  E  R  E  W  D  R  E
Query                 253   CCTTCTGTTGATTCTGCTGACAATCTATCTGAAAAATTGGAAAGAGAATGGGATAGAGAG  312
                             |||||  ||| ||||||||||||| ||||||||||||||||||||||||| |||||
Sbjct                 121   TCTTCTTCTGACTCTGCTGACAATCTGTCTGAAAAATTGGAAAGAGAATGGGACAGAGAA  180
CDS:cystic fibrosis   41     S  S  S  D  S  A  D  N  L  S  E  K  L  E  R  E  W  D  R  E
```

(c)

```
HWI-ST227:232:C1CNPACXX:7:1204:3014:160183         163
chr13     19029926  37        75M       =         19029968  117
TCGGCCATGCATGTTGTTGTCTCATGCCTGTAATCACAGCACTTTGGGAGGCCGAGGTGGGGGGTTCACCTGAAG
B@CDFDFFBFHHHIJIDHH@HEDCHHEHHJIGEHECGGGA?FAHEDC@GHGHGG5;;E=?###############
XT:A:U    NM:i:2    SM:i:37   AM:i:0    X0:i:1    X1:i:0    XM:i:2    XO:i:0    XG:i:0    MD:Z:35C28A10       ZD:Z:2.1
```

Figure 14: Examples of bioinformatics file formats. The FASTA format (a) is strictly sequence (amino acid or nucleic acid), preceded by a line starting with '>', with the name of that sequence or other identifiers. The BLAST alignment software produces more than one type of pairwise alignment output format, but shown in (b) is an example of one of the more human-readble and less machine-readble formats. These were particularly useful before web servers became more common for analysis. The SAM format (c) is the standard format for high-throughput sequencing data, particularly short read data. It is designed to be very compact, using a very specific set of codes to describe the read and its associated metadata.

ized. What we know for sure is that it is packaged in a way that it can unfold itself easily and knots and other entanglements are unlikely to form. Having a high-resolution set of contacts allows superstructures of chromatin to be modeled. (2) Enhancers are regions of DNA that bind proteins that enhance the transcription of genes, while being positioned quite far from those genes, and can even be on a different chromosome. Having a set of contacts allows the rapid discovery of previously unknown enhancers. The chromosome capture methods including 3C [172], 5C [173], and Hi-C [174] all begin by crosslinking chromatin, digesting with a restriction enzyme, then allowing ends of interacting regions to ligate in dilute conditions, forming a fusion

of two non-contiguous regions. The crosslinks are reversed and the result is subjected to further experimentation, depending on the protocol. Figure 12 from [175]) shows an overview of the various methods. One major result from the Hi-C results has been the construction of so-called "topologically associating" domains [16]. These domains generally range from one to a few megabases in size and are defined as containers for local interactions. Although a region of chromatin may interact with chromatin outside its TAD, the majority of the time it will interact with other regions in the same TAD. The methods continue to evolve but still suffer from an overall low resolution, particularly the nonspecific methods like Hi-C. Because the space of interactions is approximately quadratic to the size of the genome, current sequencing and computer technology struggle to accommodate it.

Bioinformatics Software

Bioinformatics data is stored in a variety of ways. Among the most primitive is the FASTA format [176]. Designed in 1985 for the storage of protein sequence, the files merely consist of the characters forming the sequences, preceded by the identification names of the sequences and/or other information stored in header lines (Figure 13a). The study of evolution invariably arrives at the comparison of DNA or protein sequence from different species. Among the original tools to align sequence data was BLAST [177], and with it came the need to store alignments. The original BLAST alignment file format Figure 13b is an example of a very human-readable format: allowing one to quickly compare both sequences being aligned such that alignment mismatches as well as gaps and deletions are highlighted. The alignments are each preceded by lines summarizing statistics of the alignment, again in a way that is straightforward and clear. Nearly two decades later the necessity of storing sequence alignments remains, although now priorities have shifted to making the alignment as terse as possible. The most popular format for storing the alignments from next-generation sequencing data is BAM [178], which is the compressed and indexable form of SAM (Figure 13c).

Included in the general set of expectations when submitting a genomics article to a journal, is the requirement that primary sequence or other

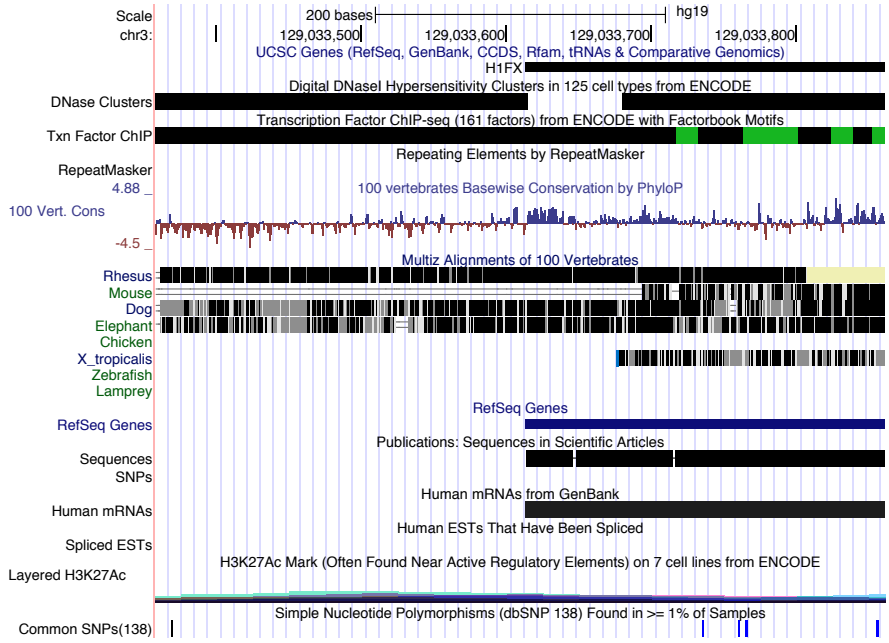


Figure 15: Example of a genome browser (in this case the UCSC Genome Browser) with multiple annotation "tracks" all aligned at a single locus for visual analysis. In this case, the locus is the region surrounding the transcription start site of the gene encoding the human histone H1 variant H1x. Other tracks visible in this case include the annotation of the evolutionary conservation of the DNA sequence (PhyloP [218], Multiz [219]), and experimental tracks including mRNA sequencing, and ENCODE epigenetic tracks [220].

large-scale datasets used in the study will be deposited in a public repository. GenBank [179], now hosted by the National Center for Biotechnology Information (NCBI), has been in operation since 1982 as a database for storing nucleotide and protein sequence and is still one of the main repositories of sequences, however it now focuses more on storing sequence for the purposes of genome assembly. Other databases such as GEO [180, 181], ArrayExpress [182, 183], the Stanford Microarray Database [184], have arisen to store data from microarray experiments. Although microarrays can be used for genotyping (SNP detection [185], copy number variations [186, 187]), alternative splicing [188, 189] or fusion gene detection [190], the majority of microarrays are used to study gene expression. Recent advancements in sequencing technology, commonly referred to as "next-generation sequencing" or "high-throughput sequencing" [191] have enabled the sequencing of

RNA-derived cDNA [192] that was previously used for microarrays. As such, GEO and some of the other databases have in turn adapted to allow the submission of sequence data as well. GEO has even gone as far as allowing not only the inclusion of the raw, primary data, but also the corresponding processed data. This processed data may take the form of continuous signal data e.g. the genome-wide depth of sequenced and aligned reads from a ChIP-seq experiment, or the RPKM/FPKM values from an RNA-seq experiment [193, 194]. It is even possible to submit further-processed data from "peak-finding" software such as MACS [195], FindPeaks [196], F-Seq [197], HOMER [198], or Pyicos [199]. The "peak" files describe positions in the genome where the signal in question is statistically significant compared to a background. Overall, there are three main ways that genomics data are used: (1) the data is mined and summarized through plots and statistics. (2) Data can be used to create or strengthen existing mathematical models that serve to simulate biological systems. (3) Data can be visualized directly through "genome browsers".

Genome browsers provide a way to visualize a variety of multiple datasets simultaneously (Figure 14). These datasets can be different types of data: RNA-seq, genes, GC percent, or even the DNA sequence itself, all aligned at the same locus. Genome browsers are typically set up in a way that the biologist is allowed to choose from a number of built-in "tracks", but also provide their own data in the form of "custom tracks". In the case of the UCSC Genome Browser, uploading genome-wide continuous data in the "wiggle" custom track format is a tiresome process: the files can be quite large and take a long time to upload. To address this problem, the bigWig format was created to allow large custom track files to reside on the biologist's side of the Internet, while being displayed on the Genome Browser side. Using a built-in index into the file, only data relevant to the region displayed on the browser is uploaded. This fixes two problems: (1) the impracticality for the biologist of using custom wiggle tracks, and (2) the massive savings in disk space for the Genome Browser website. And bearing in mind that bigWig files are compressed, for the biologist as well.

Further decentralization of UCSC Genome Browser has arrived with

the innovation of "track hubs" [200]. Instead of a single custom track, track hubs allow whole sets of tracks to be hosted external to the UCSC Genome Browser. The most notable of these are made public in the "Public Hubs" section, and include hubs from major consortia such as Blueprint Epigenomics [201], or Roadmap Epigenomics [202, 203].

While genome browsers have become indispensible tools for biologists, their purpose is still a bit one-dimensional and superficial. For publishing and communicating strong results, genome-wide datasets are most useful when they are used to produce plots, correlations and other statistics, or when they serve as the basis for mathematical models. Aside from a few crude options on genome browsers, or using a website like Galaxy [204], most data analysis is done by a bioinformatician or a group of bioinformaticians with local access to the data. It is now common for large computer clusters to perform these analyses using either custom or standard bioinformatics software. Many times these clusters are genuinely needed because highly optimized software can still be intractably slow with a large amount of data. Sometimes though, the clusters are simply used to make poorly written software usable. Software that require SAM files and are incompatible with BAM files, have a severe disadvantage to those that can read BAM files. Because the core software for reading and writing BAM files uses the C programming language, software APIs (application programming interfaces) have been written for other languages allowing many software written in many languages to be compatible with BAM. A similar situation presents itself with the bigWig format. Also, because bigWig is strictly for numerical data, it is the end product. With a fast-enough data generation pipeline, the storage of anything but the original primary data and the final computation is unnecessary. It seems rather unfortunate, that, with a file format such as bigWig, which has been in existence since 2008, very few bioinformatics software actually takes advantage of it.

Objectives

The primary objective of my thesis has been to investigate nucleosome positioning and the linker histone H1 and its variants using the dynamic chromatin system provided by inducing human T47D cancer cells with progesterone. Specific aims include:

1. Establishing genome-wide maps of nucleosomes positioned using Micrococcal nuclease (MNase), finding the well-positioned nucleosomes, and cataloging changes seen before and after treatment of progesterone.
2. Creating genome-wide maps of histone H1 variant binding, and describing the changes occurring upon induction with progesterone.
3. Finding new links between histone H1 and nucleosome positioning.
4. Finding new influences to nucleosome positioning from the underlying DNA sequence, as well as from local DNA and chromatin structure.
5. Contributing to the scientific community novel methods or software established in the course of the thesis work.

Resultats I:

The Nucleosome "Zero" in Human Cells

Andy Pohl, Guillermo P. Vicent, Ana Silvina Nacht, Giancarlo Castellano, Laura Gaveglia, Roser Zaurin, Jofre Font-Mateu, and Miguel Beato.

(manuscript)

Abstract

The nucleosomal organization of the eukaryotic genome is usually analyzed by digestion of chromatin with Micrococcal Nuclease (MNase), which preferentially cleaves the linker DNA connecting nucleosomes. In most studies the digestion is extended until the majority of the DNA products separated by gel electrophoresis accumulate in fragments of around 147 nucleotides, corresponding to the size of DNA wrapped around the nucleosomal core particle. Under these conditions genome mapping reveals a region spanning roughly -150 bp to +100 bp surrounding the transcription start site (TSS) of most genes, which seems to be depleted of nucleosomes and has been called "nucleosome-free region" (NFR). We have mapped nucleosomes in human cells using a gentler MNase digestion that generates a ladder of DNA fragments with mononucleosomal DNA fragments larger than 147 base pairs long and a dominant dinucleosomal band. Under these conditions, we detect a weak nucleosomal signal over the NFR, which is even better visible after knocking-down BRG1 and BRM, the ATPases of the SWI/SNF chromatin-remodeling complex. Moreover a digestion of free human DNA with MNase detects a valley around the TSS, indicating that the nucleotide sequence in this region is particularly sensitive to MNase cleavage. Correction for this increased nuclease sensitivity of DNA reveals a clear nucleosome peak over the TSS.. We conclude that the so-called NFR that marks the start of protein-coding genes encompasses nuclease sensitive unsusual DNA sequence occupied by a nucleosome that requires SWI/SNF for its basal dynamics.

Introduction

The organization of eukaryotic DNA in chromatin is critical for the accessibility of the DNA regulatory information recognized by transcription factors. In particular it is accepted that with the exception of pioneer factors, most other transcription factors cannot interact with DNA wrapped around the nucleosomal core particle. However, examples for factors that bind preferentially to nucleosome organized sequences have been reported [88,146]. Although this is still a debated question, it is widely accepted that the transcription start sites (TSSs) of protein coding genes are depleted of nucleosomes, giving rise to the so-called nucleosome free region (NFR) [116,117,119-121]. It is assumed that the NRF organizes the flanking nucleosomes -1 and +1, generating a regular nucleosome pattern around the start of transcription.

In conflict with this view, one of the first mammalian nucleosomes to be found well positioned occupies a region in the promoter nucleosome of the integrated Mouse Mammary Tumor Virus that partly overlaps with what should be a NFR [68,220]. This nucleosome is essential for the functional cooperation of transcription factors that is required for its hormonal induction [88,221]. Upon hormone induction, this nucleosome is remodeled by ATP-dependent complexes leading to displacement of linker histone H1 and H2A/H2B dimers [222,223] and facilitating the interaction of transcription factors with the underlying sequences assembled around a histone H3/H4 tetramer [224].

In our attempt to explore the generality of this mechanism for hormonal gene induction, we have looked at the general organization of nucleosomes in breast cancer cells and found that the hormone receptors interact preferentially with nucleosome-organized target sequences, that become remodeled similarly to the MMTV promoter nucleosome upon hormone addition [146]. For mapping nucleosome density in these cells we used relatively mild MNase digestion conditions of chromatin and performed MNase digestion of free DNA as a control followed by massive next generation sequencing. We found that the free DNA over the TSS is particularly sensitive to DNA and that upon correction for this enhanced sensitivity a clear nucle-

osome is found over the TSS, which particularly clear upon depletion of the SWI/SNF complex. Thus the concept of the NFR should be revisited as this "zero" nucleosome may play important roles in gene regulation.

Materials and Methods

Cell culture and progesterone treatment

Cells were cultured in RPMI 1640 medium supplemented with 10% FBS, 2 mM L-glutamine, 100 U/ml penicillin and 100 μg/mL streptomycin at 37°C in a 5% CO_2 containing atmosphere. Cells were seeded at 25% confluence and cultured in RMPI medium, without phenol red and supplemented with 10% dextran-coated charcoal-treated FBS (DCC/FBS). 48 hours after seeding the cells, the medium was replaced by fresh RMPI medium without FBS. 16 hours later, cells were incubated at 37°C, with the progesterone analogue R5020 (both at 10 nM) for the time points indicated at each experiment.

MNase digestion *in vivo*

It is recommended to test a range of nuclease concentrations when working with a new cell line, but in this case, we used 2.5-3.5 X 10^6 T47D cells grown on 10 cm Petri dishes (approx to 70% confluence). After washing the cells with 10 ml 1X PBS at 37°C, the cells were covered with 2 ml of Buffer A at 37°C, supplemented with 0.5 mg/ml lysolecithin, let to stand for 1 min at 37°C, afterward removing the buffer. The cells were covered with 2 ml of buffer A at 37°C containing MNase. Prior to the final experiment, MNase was titrated in buffer A with the following concentrations: 0, 30, 90, 270, and 800 U/ml. The cells were incubated 2 min at 37°C and the reaction was stopped with 160 ml of stop solution (40 mM, final concentration). The cells were then scraped, and collected in a 15 ml falcon tube, then centrifuged at 3400 x g, 4°C, for 2 min. The resulting pellet was washed with 2 ml of cold 1X PBS, and centrifuged again, afterwards discarding the supernatant. The pellet was then resuspended in 600 ml of Lysis buffer II and incubated for

10 min on ice, followed with the addition of 7 ml of 10 mg/ml RNase A and incubated 30 min at 37°C. The RNase treatment was followed by the addition of 50 ml of 10 mg/ml Proteinase K and incubated for one hour at 45°C. Digested DNA was then extracted using phenol-chlorophorm, resuspended in 1X TE, and had its concentration measured using a Nanodrop ND-1000 spectrophotometer. Finally, the digestion pattern was verified on a 1% agarose gel, with the mononucleosome or dinucleosome band excised and gel-purified. Resulting DNA fragments were then used in the preparation of Illumina paired-end sequencing libraries.

Buffers

- MNase solution (Worthington, Lakewood, NJ). Dissolve at 45.000 U/ml in BSA 0.1%. Store the stock solution at -80°C in small aliquots and use only once.
- Buffer A: (filtered, -20°C): 15 mM Tris-HCl, pH, 7.5; 150 mM sucrose; 15 mM NaCl; 60 mM KCl; 2 mM $CaCl_2$; 0.15 mM spermine (added just before use); 0.5 mM spermidine (added just before use).
- Lysolecithin (L5254, SIGMA, St. Louis, MO) in buffer A at 0.5 mg/ml (prepared just before use).
- Stop solution: 500 mM EDTA, pH 8.0.
- Lysis buffer II: 50 mM Tris-HCl , pH 8.1; 1% SDS, 10 mM EDTA.

Microarray, RNA-seq and ChIP-seq

RNA-sequencing was performed by collecting purifying total RNA from untreated cells and cells treated with progesterone for 6 hours. Ribosomal RNA was depleted using the Ribominus kit (Invitrogen), then remaining RNA was used to construct Illumina paired-end sequencing libraries, then sequenced 2 x 50 nucleotides. Sets of genes for all of the analysis were selected based on the RNA-seq (details for the construction of these gene sets may be found in the Results II chapter's methods section). For the knockdown Brg1/Brm

cells, microarrays were performed using the Agilent platform, in triplicate, against RNA from cells treated with an siRNA control in a competitive hybridization.

ChIP-sequencing was performed following the protocol mentioned in the methods of the Results II chapter, using antibodies for H2A/H4 (gift from Dr. D.S. Dimitrov), and H2A.Z (Abcam ab4174).

DNase-seq was performed using the protocol from Song, *et al.* (2010) [166], with modifications of the protocol to accommodate T47D cells.

Sequence data processing

Each sample was sequenced in its own flow cell lane using an Illumina HiSeq 2000. FASTQ files were aligned to the reference human genome GRCh37/hg19 using BWA [225]. At the time of writing, BWA does not fill all BAM fields completely, so "samtools fixmate" was run to fill in information about a paired-end read's mate. BAM files were tagged and/or flagged according to the following criteria:

1. If 40% or more of a read's quality values were demarked by a '#' (Illumina 1.8 Phred quality), meaning the quality value at that base is unknown, then the read is flagged with the 0x200 bit (meaning the read does not pass quality controls), and tagged with ZL:Z:0.40.

2. If the read overlaps with a highly-duplicated region (HDR) [226], the read is also flagged with the 0x200 bit and tagged with "ZR:Z:HDR".

3. If both the read and its mate's coordinates match one or more pair's coordinates, the read is tagged "ZD:Z:x.y", where in this case x indicates the total number of pairs duplicated, and y represents an index for one of the pairs in question, numbered from 1-x. Reads with y > 1 are also flagged with the BAM flag 0x400 bit set, which is the standard flag for PCR or optical duplicates.

4. If flags and tags are encountered on a read's mate, then they are added to the read.

Unmapped reads, and mapped reads with unmapped ends were retained in the BAM files but not used in subsequent analysis.

Sequence depth profiles

Fragment depth profiles for each sample were created using custom software that provides the equivalent functionality of "samtools depth" or "bedtools genomecov". These depth profiles, normalized in several ways, provide the basis of the remaining analysis.

Karyotype normalization

Accommodating the polyploidity of the T47D genome was done by either of two methods:

1. We manually constructed a genome-wide karyotype of the T47D cell by combining information from the following:
    a. Visual inspection of genome-wide raw read depths from various sequencing experiments.
    b. Spectral karyotypes of T47D created using M-FISH [227].
    c. Inter-chromosomal interaction matrices from a Hi-C experiment (Le Dily et al, 2014) [248].
2. Computing local mean depths. This approach requires a parameter w for the size of the window to compute the mean locally. The equation is then:

$$LM(i) = \frac{1}{w} \sum_{j=i-w/2}^{i+w/2} D(j)$$

of values of i in the range:

$$w/2 < i < c - w/2$$

where c is the size of the current interval (chromosome). For our purposes we chose w to be 20000, which we thought to be sufficiently large, being roughly the size of 100 nucleosomes (including linkers). In order to avoid zero-denominator problems using this term later in scaling ra-

tios, we removed all regions where the mean depth was less than 0.1 for 20kb.

Genome-wide, the effect of normalizing the karyotype can be seen in Figure S2.

Depth profile normalization

Each sample was sequenced at a slightly different depth; therefore the resulting genome-wide read depth profiles were scaled to the mean depth (total fragment bases divided by the total coverage) of the input sample. We can define depth at a base *i* as *D(i)*. The mean is simply:

$$M = \frac{1}{n}\sum_{i=1}^{n} D(i)$$

where n the number of **used** bases in the genome. A common strategy in normalizing between samples is to scale using ratios having some constant in the numerator by each sample's total number of reads/fragments. In the case of paired-end sequencing, minor differences in sonicated fragment lengths can contribute to differences in read depth. Using a single scaling constant genome-wide is insufficient to represent the T47D cell's polyploidy karyotype while also using the reference genome for read mapping and browser visualization. In order to counter effects of chromosomal duplications and transposons, we also utilize a background input DNA's overall mean depth over its local mean depth across a 20 kb window. The scaled depth is then:

$$NormD(i) = \left(\frac{M_{inp}}{M_{sam}}\right)\left(\frac{M_{inp}}{LM_{inp}(i)}\right) D_{sam}(i)$$

The naked MNase-seq sample depth was then subtracted from each sample depth. The final quantity was then multiplied by the 50 bp CRG/ENCODE mappability: a value between 0 and 1 designed to suppress mappability artifacts arising from repetitive DNA [228] (see Figure S3). The normalized, naked MNase-subtracted sample occupancy is calculated as:

$$InpSub_{sam}(i) = \left[NormD_{sam}(i) - NormD_{inp}(i)\right] MAP(i)$$

For the Yoruba fibroblast MNase-seq samples (GM cells), we instead used the local/global means of the MNase sample itself, instead of naked DNA. Because this is used for correcting larger-scale anomalies, this kind of self-normalization is still suitable.

Nucleosome positioning score

To calculate the nucleosome positioning score, we used an algorithm similar to the dyad positioning score from Gaffney, *et al* (2012) [79]. The algorithm begins by first finding all of the fragment midpoints, then simply calculating a ratio (at each base in the genome) of the number of fragment midpoints within 15 bases over the number of midpoints within 100 bases. An array of well-positioned nucleosomes will appear as successive sharp peaks, centered at a consensus nucleosome position. Although it is possible to refine this method further: with the assumption that the two windowing parameters can vary depending on cell line-specific nucleosome spacing, in practice the method seems to be quite discriminatory without further optimization. For the positioning score for dinucleosome centers, we used the same scoring algorithm, except using 375 bases instead of 200 for the large window, and 51 bases instead of 15 for the smaller window.

Clustering nucleosomes

To find super-occupied NFR genes, we used the k-means clustering functionality of bwtool [229], first on the whole set of genes (14,561). One gene (NBPF1), had a particularly low background (20 kb local mean) and was removed. Other than this gene, the clustering algorithm requires a set of data without missing values, so the software discarded 63 more genes that most likely in regions where the digested free DNA background was extremely low. We used k=2, k=3, k=4, k=5, and k=6, and visually inspected the resulting clusters by plotting the average profile. Figure S4a-e shows each of these plots, and while the larger cluster appeared closer in appearance to the canonical MNase NFR, the smaller one retained the trend that the whole set had, albeit with higher occupancy. We then ran the clustering again with k=3 and k=4. Using a k=3 yielded two nucleosome-depleted clusters, with a third

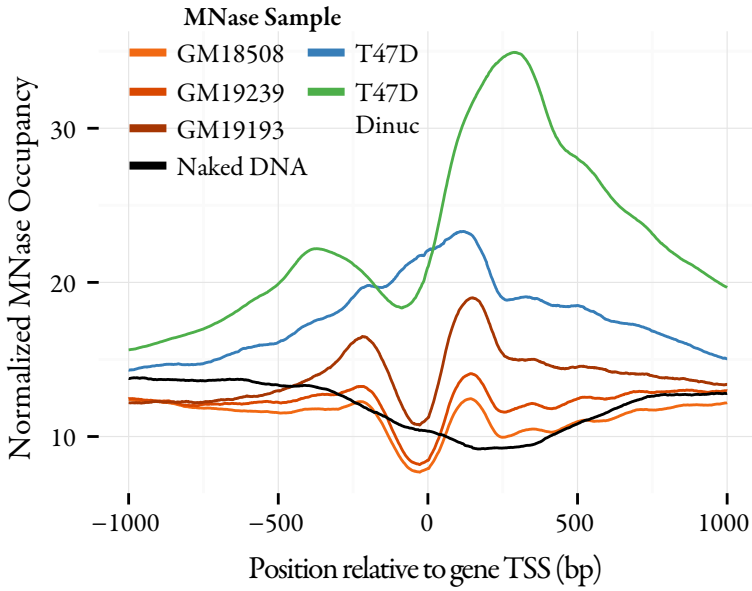


Figure 1: Aggregate profiles of MNase samples at the TSS of 14,561 genes. The Yoruba fibroblast samples are from Gaffney, *et. al* 2012 [79], and display a strong nucleosome-free region (NFR). Our T47D mononucleosome sample (blue) has no such depletion in this region, though it does show closely-similar peaks at the -1 and +1 nucleosomes to the GM samples. Our T47D dinucleosomes sample however, does have a depleted region.

visually distinct form the other two. Upon using k=4 we found that this third cluster was essentially split in two, so for the remaining analysis decided to focus on two groups: (1) the combination the two NFR clusters from the k=3 clustering (the "normal" set of genes), and (2) the third cluster in that same clustering set (the "super-occupied nucleosome zero" set of genes).

Results

The DNA from MNase experiments was size-selected by electrophoresis. To ensure the cells were indeed subjected to the light dosage of MNase we had intended, we checked that the dinucleosomal band had a higher molecular weight than the mononucleosomal band before isolating the DNA from the gel (Figure S1). After sequencing, and some quality controls including removing duplicates fragments, we were left with the amounts of sequence seen in Table S2.

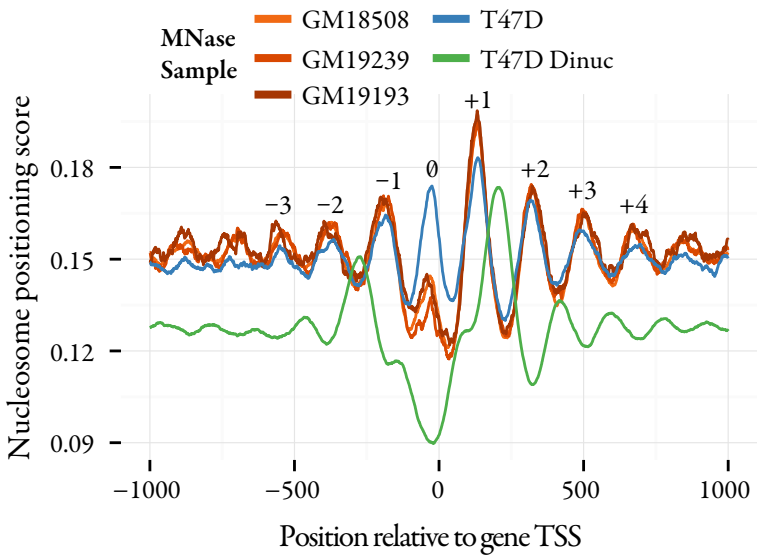


Figure 2: Nucleosome positioning score calculated all samples other than the dinucleosome, show a positioned nucleosome just upstream of the TSS. The GM samples, which showed a distinct NFR region, still show a nucleosome where our T47D "zero" nucleosome is located, albeit with weaker positioning than our T47D. The dinucleosome center positions overlap in the anti-phase of mononucleosome dyads, and have a tendency to not include the zero nucleosome.

Fragment depth profiles were generated normalized for our T47D samples and compared to profiles of MNase-seq experiments from Yoruba fibroblasts GM18508, GM19193, and GM19239 at the TSS of 14,561 genes we chose, discarding those with complex loci or very low RNA-seq expression (Figure 1). Having been treated with a stronger MNase digestion, the fibroblast samples display the canonical NFR, while our T47D lacks this depleted occupancy. As we mentioned, we choose a level of MNase digestion where the dinucleosome band on the MNase ladder carries a higher molecular weight than the mononucleosome band. It is interesting then that this sample shows depletion in the NFR region compared to its own occupancy in the +1/+2 and -1/-2 nucleosomal regions. Overall, both upstream and downstream of the TSS exists a depletion of occupancy in the MNase-treated free DNA, or "naked" DNA sample, indicating the region is more sensitive to MNase.

We calculated genome-wide positioning scores using a slightly-modi-

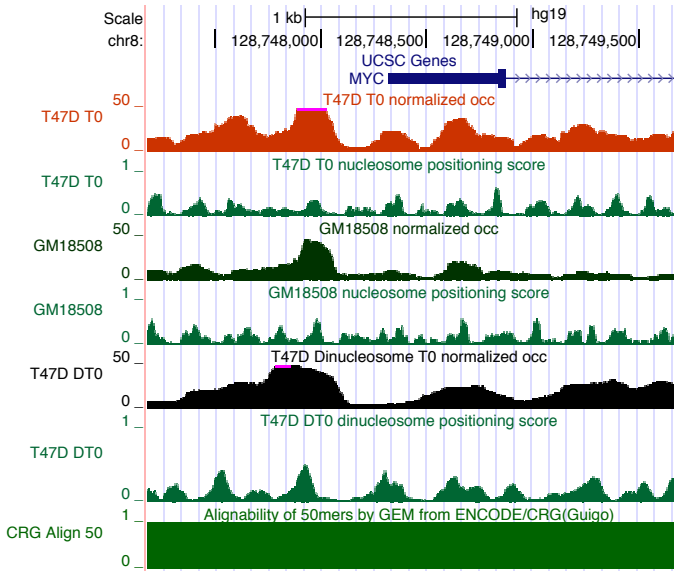


Figure 3: Genome browser screenshot showing a "super-occupied nucleosome zero" at the TSS of MYC, in the T47D track. Reasonably-regular phasing of the nucleosomes is seen in the track just below in the nucleosome positioning score. For comparison, the GM18508 MNase-seq occupancy is also shown, and illustrates a somewhat-depleted NFR region between the -1 and +1 nucleosomes. Finally, the mappability track is also shown to demonstrate this lies in a region not known to produce artifacts of short read alignments.

fied version of the algorithm proposed in Gaffney, *et al* 2012 [79]. The score is calculated as a ratio of the number of nucleosome dyads (paired-end fragment midpoints) falling within 15 bp of a locus over the number of dyads within 100 bp. After obtaining these scores, a clear nucleosome appears in our samples, centered roughly -25 bp from the TSS (Figure 2). This nucleosome appears in the Yoruba fibroblast samples as well, the only difference being a lower positioning score, probably owing to the lowered occupancy compared to the flanking regions. In all samples, we observed the highest positioning scores in the +1 nucleosome. On this basis, we also aligned occupancy and positioning to the high-scoring bases instead of the TSS (Figure S4), and were able to see the zero nucleosome in the T47D samples a bit better. We also calculated a positioning score for the dinucleosomal fragments, with the fragment midpoints being interpreted as the center of a nucleosome linker. We adjusted the window size parameters to accommodate the larger fragments, but in lieu of optimizing these parameters we were left with

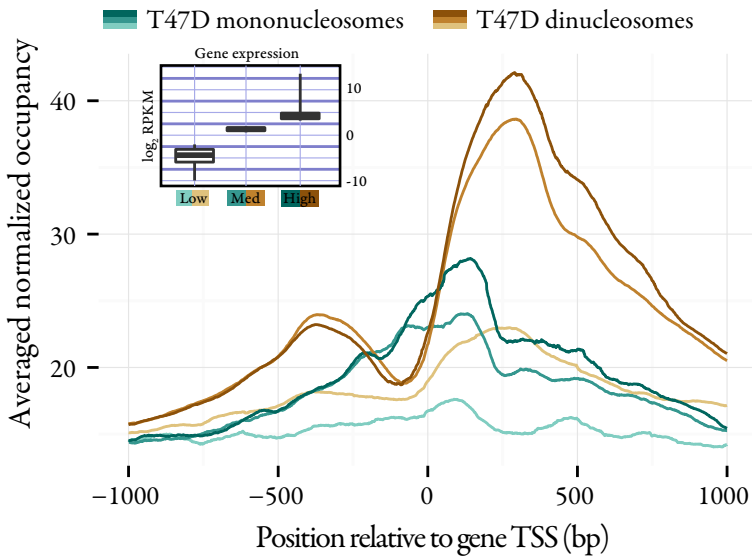


Figure 4: Aggregate profiles of mononucleosome and dinucleosome MNase samples at the TSS of genes at three different levels of expression (see inset figure). Genes with higher expression tend to also have higher nucleosome occupancy, particularly in the +1 nucleosome.

somewhat lower scores in general, also because the dinucleosome phasing at the TSS, apart from the zero nucleosome, displays overlapping, and anti-correlates well with the mononucleosomal phasing. Figure 3 shows an example of the nucleosome positioning and occupancy of the TSS of the MYC gene in the UCSC Genome Browser. The example highlights some of the basic trends seen genome-wide, but in this case shows a higher positioning of the dinucleosome center at the nucleosome zero locus than is typical.

In order to gauge the association between transcription and nucleosome occupancy, we sorted genes by their RNA-seq RPKM, divided them into five equally sized groups, and then used the three groups with the lowest, the middle, and the highest expression to make Figure 4. Higher nucleosome occupancy occurs in genes with higher expression, particularly in the dinucleosomal sample. Nucleosome positioning (Figure S6) at different expression levels also reveals that at low expression levels, nucleosomes are poorly positioned, but once expressed there is similar positioning, particularly at the zero nucleosome. We estimate that in the genes with the highest level of expression, there is a shift of 7-10 bp with the -1 and +1 nucleosomes away from

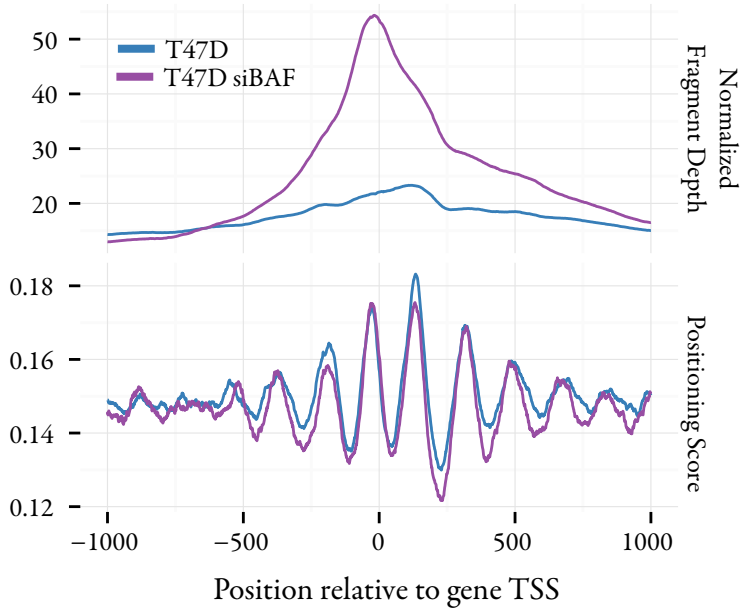


Figure 5: Occupancy (upper plot) and positioning (lower plot) of mononucleosomal samples at the TSS of 14,561 genes, with and without a knockdown of the BAF complex (SWI/SNF). We observe a large increase in occupancy across the region 500 bp upstream and 1,000 bp downstream of the TSS, peaking at the proposed zero nucleosome locus: 25 bp upstream of the TSS. Positioning of the nucleosomes is not very different between the samples when considering the full set of genes.

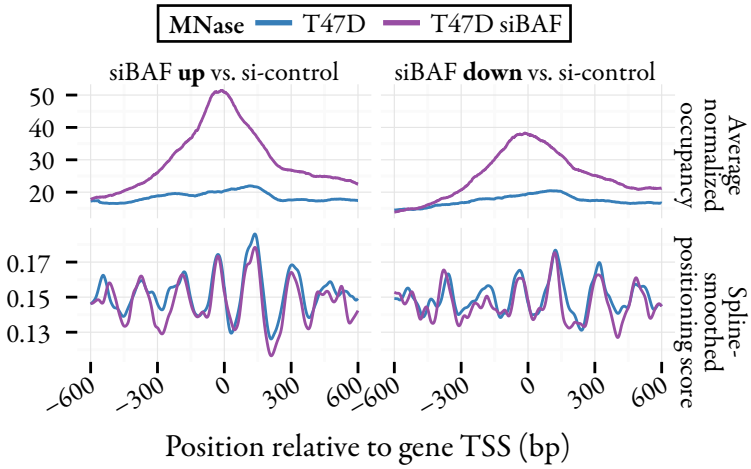


Figure 6: At genes affected by the knockdown of the BAF complex: 1,266 genes had higher expression (FC ≥ 1.4) than the si control, while 1,178 genes had lower expression ($FC^{-1} \leq 1.4$).

the zero nucleosome, compared with the mid-expressed genes.

To investigate the role of the SWI/SNF-like BAF chromatin remodeling complex, we treated T47D cells with siRNA targeting the Brg1 and Brm genes, and performed MNase-seq experiments using these cells. Overall, nucleosome occupancy in the NFR region increases quite dramatically, while nucleosome positioning does not change (Figure 5). We performed microarrays using the competitive hybridization of the mRNA-derived cDNA from siBAF-treated cells versus those with an siRNA control to find genes dependent on Brg1/Brm1 ATPase activity. We found 1,266 genes with an expression fold change of 1.4 or greater in the knockdown sample versus the siRNA control, and 1,178 genes with an opposite change in expression favoring the siRNA control. We examined the nucleosome occupancy and positioning in both MNase-seq datasets for both sets of the siBAF-regulated genes (Figure 6), but did not see any major difference between the MNase experiments that we did not already observe when looking at the full set of genes.

Treating cells with progesterone induces a response involving the up-regulation and down-regulation of various genes and widespread chromatin remodeling [146]. We also performed MNase-sequencing with both BAF-knockdown and normal T47D cells. In both cases, after progesterone induction, we see sequenced nucleosomal fragments map to a larger percentage of the genome, and we recover fewer well-positioned nucleosomes (see Table 3 of the Results II chapter). Other changes are induced by progesterone have been reported [146], but we see here nucleosome occupancy at the TSS dominated by the presence of the zero nucleosome. At all of the genes together, we see a general effect of the nucleosome occupancy increasing after progesterone in the wild type T47D sample, but decreasing after progesterone in the BAF-knockdown sample. Positioning of the nucleosomes does not change significantly after progesterone induction in either case. In wild type T47D cells, we identified genes by RNA-sequencing as being up-regulated (1,046) or down-regulated (587) after 6 hours of progesterone treatment compared to their basal expression level. With the microarrays mentioned earlier performed on BAF knockdown cells, we also did a progesterone treatment, but in this case only finding 84 genes up-regulated, and 28

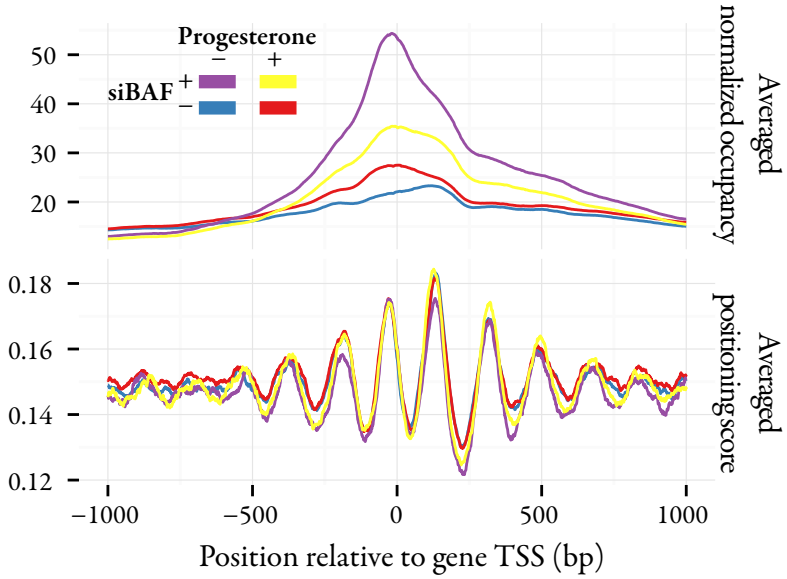


Figure 7: The effect of progesterone is much more apparent in nucleosome occupancy (upper panel), than in positioning (lower panel). Increased nucleosome occupancy surrounding the TSS is seen both when treating with BAF siRNA or with progesterone. With both treatments though, the occupancy is reduced compared to the occupancy after treating with the siBAF RNA alone.

genes down-regulated by progesterone. In Figure S8 we show nucleosome occupancy and positioning for each sample at the progesterone-regulated genes, but in all cases the trend follows the larger trend seen in all genes (Figure 7).

We performed ChIP-seq experiments using antibodies against all variants of histone H2A, the histone H2A variant H2A.Z, histone H4, and histone H1 variant H1.2 to investigate the range of histone content across the promoter. To somewhat varying extents, H1.2, H2A, and H4 all show strong depletion in the region surrounding the TSS. H2A.Z on the other hand, is highly enriched at promoters (Figure 8a). To see how well the H2A.Z binding correlates with nucleosome zero occupancy, we took the average raw read/fragment depth from 50 bp upstream of the TSS to the TSS from each of the H2A.Z T0, MNase T0, and siBAF MNase T0 datasets, at each gene. Spearman correlations for H2A.Z versus MNase, and H2A.Z versus the siBAF MNase were R = 0.406 and R = 0.411, indicating that not only are H2A.Z binding and nucleosome zero occupancy correlated, but despite different levels of occupancy, the MNase samples correlate similar-

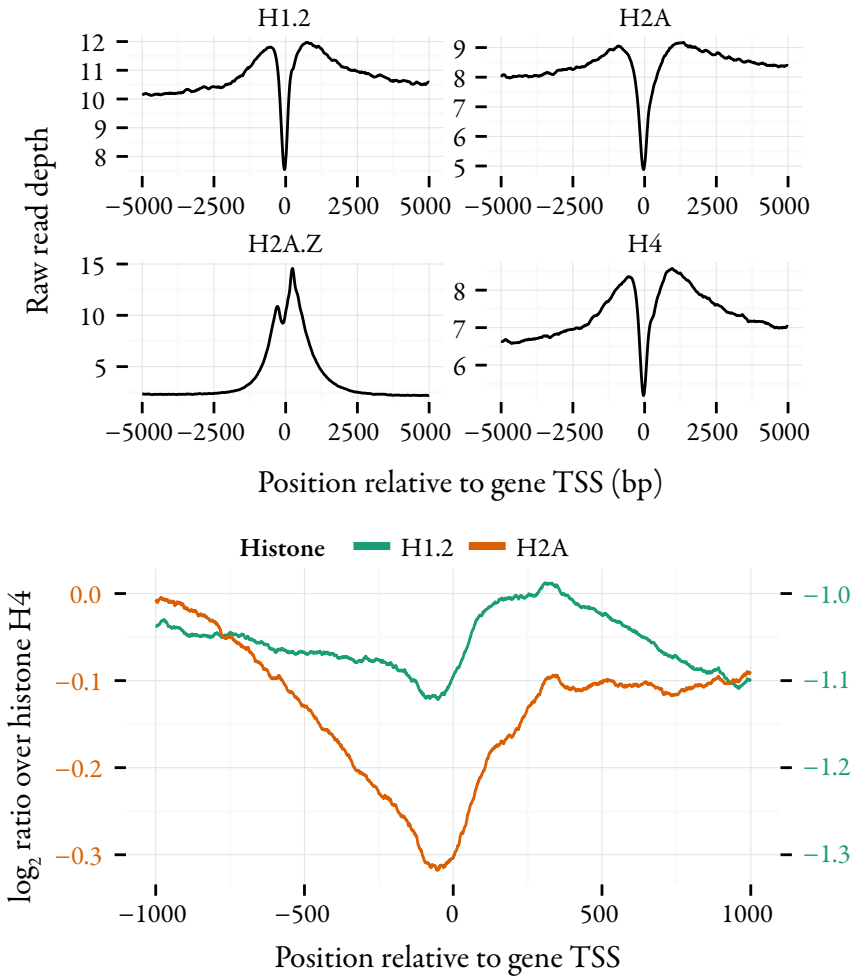


Figure 8: (a) Raw profiles of core histones at TSS. (b) Content of histones H1.2 and H2A expressed as log-ratios to H4. At the peak of the zero nucleosome region lies the greatest depletion of H1.2 and H2A versus H4.

ly with H2A.Z. In fact, between the two MNase samples, the nucleosome zero occupancy correlation is R = 0.774. As core histones, H2A and H4 are both ubiquitously deposited throughout the genome, while variants of histone H1 have some specificity but are also very widely deposited [52]. We calculated genome-wide ratios of H2A and H1.2 versus H4 (Figure 8b), with the assumption that on average H2A content will be equivalent to H4, while H1.2 will be half. H2A in particular is sharply depleted at nucleosome zero, indicating the possibility that this nucleosome is often not composed of a full

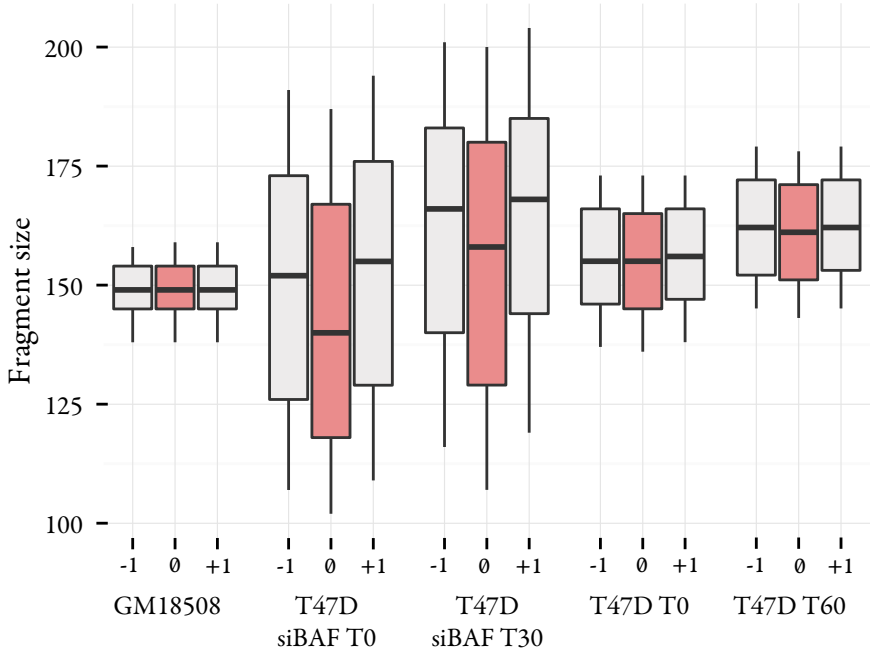


Figure 9: Paired-end fragment sizes taken from the DNA regions generally comprising the -1, +1, and the proposed zero nucleosome. Fragment sizes in the zero nucleosome (highlighted in red), tend to be smaller, particularly in the siBAF samples.

histone octamer.

To explore the possibility nucleosome zero has an abbreviated structure, we also examined the sizes of the fragments contributing to nucleosome zero's presence, and compared that distribution of sizes to those from the -1 (-300 to -150 upstream of the TSS) and +1 (+75 to +225 downstream of the TSS) flanking nucleosomes (Figure 9). As expected with a stronger MNase digestion, the fibroblast MNase-seq fragment size distribution is very tight, and unchanged between the zero nucleosome and its flanking nucleosomes. By contrast, the T47D samples have smaller fragment sizes in the nucleosome zero region, particularly in the BAF knockdown sample from cells not treated with progesterone, where we see a mean fragment size reduction of 12-13 bp.

Lastly, we looked at the other main nucleosome-free region: surrounding the TTS of genes. The nucleotide composition of the TTS regions of the genes used in our analysis was much different than that of the TSS regions. A generally GC-rich region encompasses the sequence surround-

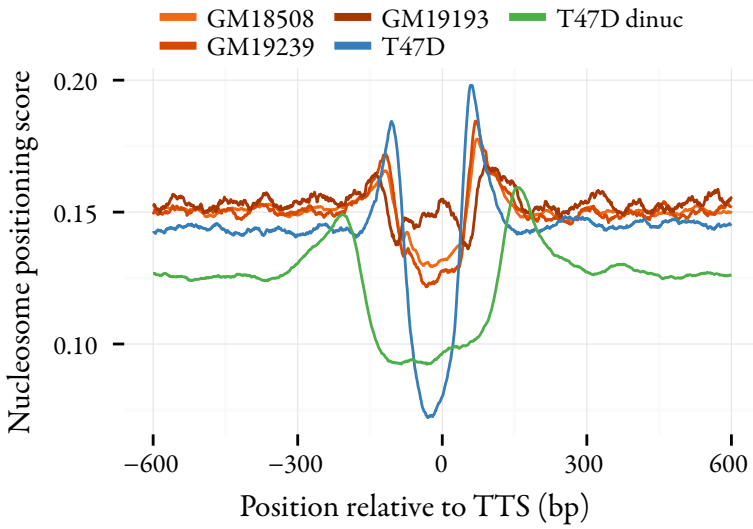


Figure 10: Nucleosome positioning at the transcription termination site. Reasonably well-positioned mononucleosomes on either side of an AT-rich spike at around 22 bp upstream of the TTS. These mononucleosomes are also seen to be included in dinucleosomes.

ing the TSS, with well-known exceptions located at sites such TATA-boxes. The TTS region contrarily is much more AT-rich, peaking at a region of 80% AT-richness 20-25 bp upstream of the TTS (Figure S9a). Due to the cleavage bias of the MNase nuclease to AT dinucleotides, the use of MNase to examine nucleosome occupancy in this region results in very similar profiles among the samples, including the MNase-digested naked DNA sample (Figure S9b). Nucleosome positioning around the TTS is quite poor compared to the TSS (Figure 10). What seems apparent is that although the TSS appears to have a region of nucleosome depletion, the distance between the -1 and +1 nucleosome is not much more than what would be expected from any two nucleosomes.

Discussion

After performing nucleosome-positioning experiments in a human cell line using MNase, we found that with a gentle digestion, additional nucleosome

occupancy is visible in the nucleosome-free region. This nucleosome is as well-positioned as its flanking nucleosomes, and is basically in phase with them. It appears in all four of our experimental datasets positioning mononucleosomes: with and without a BAF protein siRNA, and with and without progesterone. Knocking down BAF proteins increases the occupancy of the "zero" nucleosome while largely maintaining the same positioning.

Using a gentle digestion of MNase also enables us to capture the dinucleosome for positioning analysis. The case of the dinucleosome is interesting. We have seen that dinucleosomal fragments do not usually include the zero nucleosome. One interpretation is that the linker size before and after the zero nucleosome is slightly longer than normal: 50 bp versus around 43-44 bp average, without considering H1 (see phasogram Figure S7). This linker could be even longer if the zero nucleosome is lacking one of its H2A/H2B dimers. We believe that when present, the zero nucleosome contains at least the histone H3/H4 tetramer. 93% of the previously reported "fragile" nucleosomes in the nucleosome-free regions of yeast were recovered with histone H3 affinity purification, which is comparable to the 98% of total nucleosomes that were recovered [230].

Levels of H2A.Z binding increase with the increased occupancy of the zero nucleosome, indicating this to be a typical member of that nucleosome. As we believe the zero nucleosomes to be unstable, we think that these nucleosomes will often contain the histone H3 variant H3.3 as well, owing to previous observations that H2A.Z/H3 nucleosome core particles (NCPs) are as stable as H2A/H3 NCPs, but H2A.Z/H3.3 is much less stable, as is H2A paired with H3.3 [231]. The pairing of H2A.Z and H3.3 was also shown to be enriched in the NFR in HeLa cells [232], with a nucleosome stability more sensitive to salt concentration than canonical NCPs. H3.3 requires CHD1, another ATP-dependent chromatin remodeler also in the SWI/SNF, for its deposition [233]. For this reason, it may be interesting to see to what extent knocking down CHD1 might have on nucleosome zero occupation.

The zero nucleosome in the context of progesterone is a not altogether clear. We see general effects, where with either an siRNA treatment against BAF proteins or a progesterone treatment, there is an increase in nucleosome

occupancy at the zero nucleosome. But both treatments together reduce the level of occupancy seen in with the BAF knockdown alone. We do not have enough data to explain this phenomena, however we speculate it may be due to one or both of the following reasons: (a) the BAF proteins could only be knocked down to a level of around 20-30% of their expression without causing death to the cells. Perhaps the remaining Brg1/Brm proteins are able to perform their tasks more efficiently with the liganded progesterone receptor bound in the same remodeling complex. (b) The increase in occupancy after progesterone with in the wild type cells may be more related to increased transcription in genes up-regulated by progesterone (c) The progesterone treatment was shorter (30 min versus 60 min) in the siBAF sample than in the wild-type T47D. Although we think it is unlikely, perhaps the occupancy is reduced at a period of 30 min and is increased afterwards.

Other than the roles chromatin remodelers or transcription may have with the zero nucleosome, we believe that the underlying properties of the promoter enhance its occupation. The promoter has been shown to harbor sites of non-specificity for transcription factors where elsewhere the factors recognize more specific sequence [234]. Transcription factor binding sites appear in the promoter region with periodicity peaking at 10-10.5 bp and have phasing correlated with nucleosome phasing [235]. In addition, simulations of the biophysical properties and molecular dynamics of the promoter sequence compared with non-promoter sequence has provided the predictive power to successfully find novel promoter regions *ab initio* [122,236]. In the course of our analysis, we encountered a group of genes having particularly high nucleosome zero occupation, which we refer to as "super-occupied" genes. When compared to the other genes, they seem to have all of the distinctive sequence-based and structural features of the promoter, but more exaggerated. However, we could not link these genes to a meaningful hypothesis, so for the moment they remain a curiosity. They were not expressed at levels outside the typical range of all genes, and they were not enriched in an informative GO category (Table S1). But they did exhibit higher GC/GG dinucleotide content, lower AA/AT content, higher predicted hydroxyl-radical cleavage site content [237], as well as decreased roll/shift stiffness

and increased roll flexibility in terms of DNA helical deformation [236] (see Figures S10, S11). We can perhaps postulate that the heavy occupation of nucleosome zero in super-occupied genes is more of a consequence of idealized sequential/structural conditions in those promoters contributing to increased nucleosome zero stability.

The DNA sequence around the TSS has special properties and favors non-specific binding of transcription factors and nucleosomes. These nucleosomes are dynamically unstable and a preferred substrate for SWI/SNF favoring nuclease cleavage. However, in the absence of SWI/SNF the nucleosomes become clearly visible. The notion that the promoter includes a region of DNA constitutively free of nucleosomes is a convenient explanation for nucleosome positioning experiments that show a high depletion upstream and including the TSS, because it offers the immediate conclusion that the initiation of transcription has very direct accessibility to DNA. But in addition to transcription factors or polymerases, mutagens would also have increased accessibility, and it has been shown that an important role of nucleosomes is its ability to protect against mutation [18]. Although convincing work has been done in yeast and HeLa cells that nucleosome-free regions are not nucleosome-free, the idea persists. An article was recently published exemplifying perfectly the dogmatic view of NFR. In it, the authors describe a new method of positioning nucleosomes, called ATAC-seq, which uses a transposase loaded with sequencing adaptors that binds preferentially to DNA in open chromatin [238]. ATAC-seq is a method to recover mononucleosomal or polynucleosomal sized fragments in a way that combines the usefulness DNase-seq and MNase-seq. They demonstrated their method on human CD4+ blood cells, and recovered fragments corresponding to nucleosomes -2, -1, +1, +2, +3, etc. surrounding the TSS of genes. They also recovered fragments corresponding to a zero nucleosome, but simply dismiss these fragments as non-nucleosome-bound, because they are smaller than their size threshold -- completely neglecting the possibility that the fragments are associated to a partial nucleosome. Though others have shown evidence to the contrary in yeast and HeLa cells, we have extended the concept to suggest the removal of the zero nucleosome is mediated by SWI/SNF.

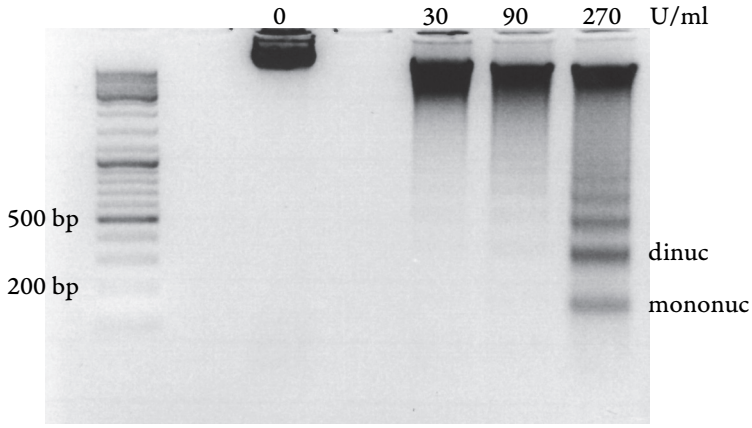


Figure S1: The *in vivo* MNase digestion ladder from untreated T47D cells showing a stronger dinucleosomal band than the monocleosomal band (far right lane). Also shown are several weaker concentrations of MNase.

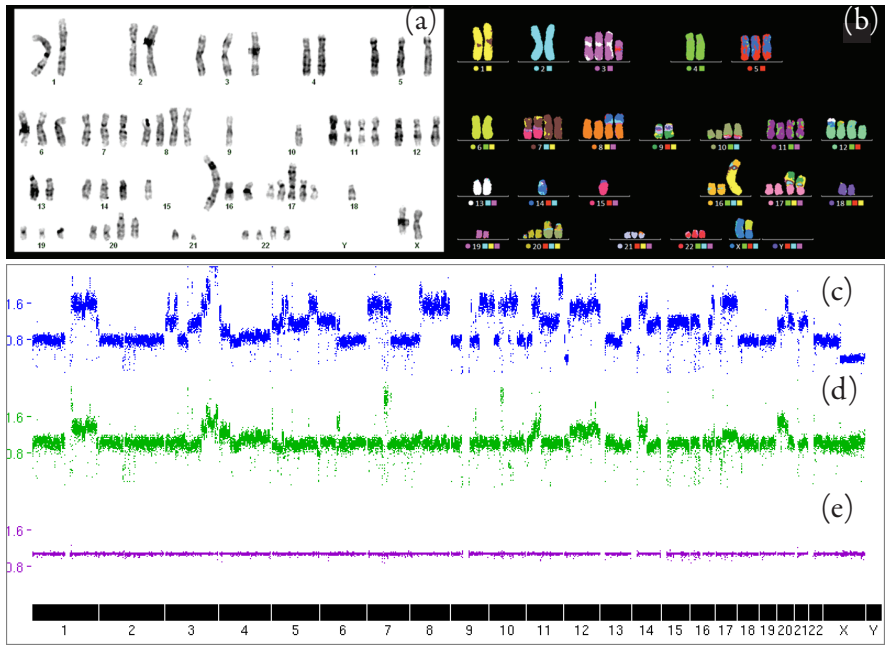


Figure S2: The polyploidity of the T47D genome seen by (a) G-banding (b) M-FISH (from Rondón-Lagos, *et al.* 2014) [227]. Much of the genome is triploid or more, while some regions e.g. chromosome X, are missing a second copy. Such a karyotype results in genomic data varying drastically in depth by chromosome, and is particularly problematic when performing an analysis such as clustering. (c) shows an unnormalized ChIP input fragment depth genome-wide. Results of normalizing (c): with the "karyotype" normalization method (d), and with the "background local mean" method (e).

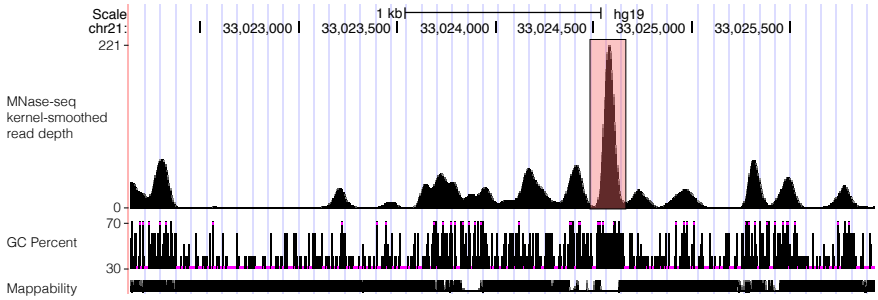


Figure S3: False positive artifact (highlighted in pink) introduced into results due to a pileup of reads mapping to a poorly-mappable region.

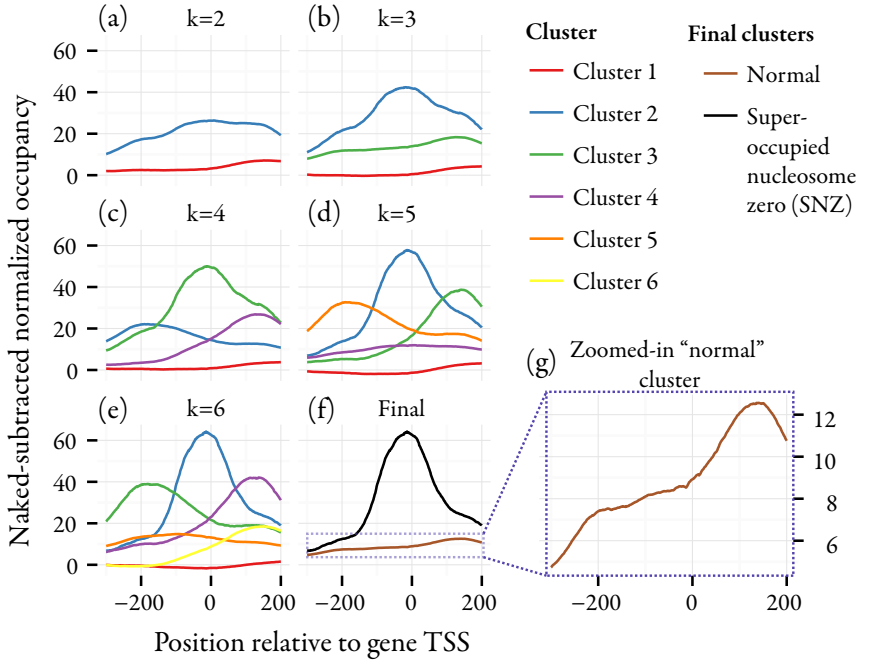


Figure S4: Clustering analysis: we progessively used a larger and large k until it seemed that all the non-SNZ clusters had redundancy (e), i.e. cluster 4 (1,384 genes) is a higher-occupancy version of cluster 6 (2,778 genes), and cluster 3 (1,002 genes) is a higher-occupancy version of cluster 5. Cluster 1 (4,690 genes) is the typical TSS profile, and it's combined with clusters 3-6 to obtiain the "normal" TSS cluster seen in (f) and (g). It should be noted that this profile does not match the canonical TSS profile for MNase-seq data. The reason for this is that the canonical profile is somewhat lost after subtracting the profile of MNase-digested free DNA.

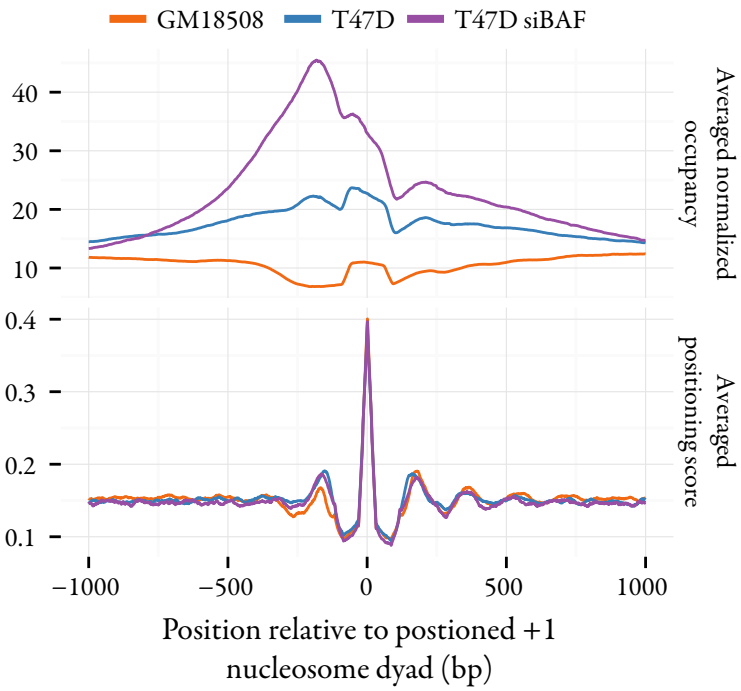


Figure S5: Occupancy and positioning profiles when aligned to base with highest positioning score 50-250 bases downstream of the TSS.

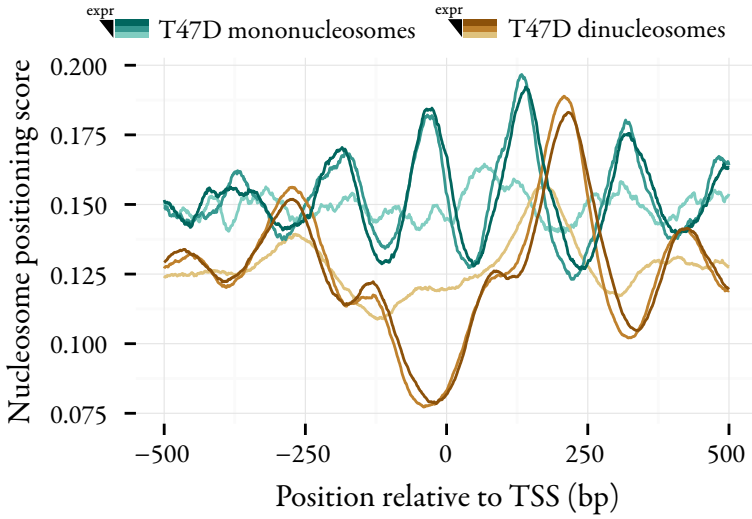


Figure S6: Nucleosome positioning at three gene expression levels.

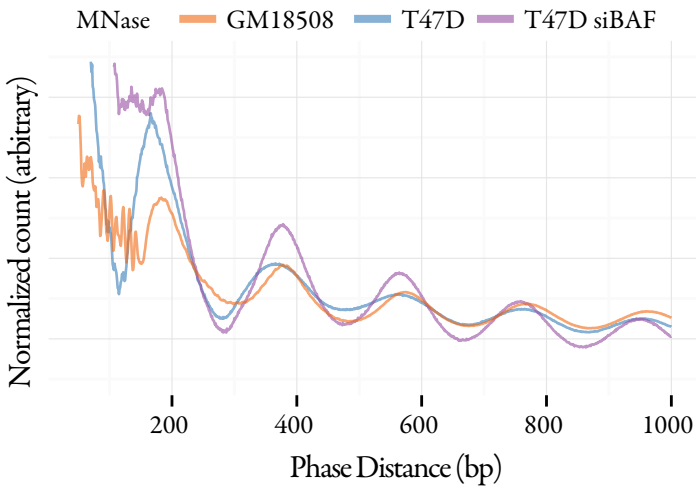


Figure S7: Phasogram of mononucleosome samples. Due to a positive skewness in the phase distributions, exact phasing is difficult to ascertain, but in a previous nucleosome positioning experiment using single-end MNase-seq, we had estimated 191 bp as the general spacing of T47D nucleosomes, both with and without treatment of progesterone.

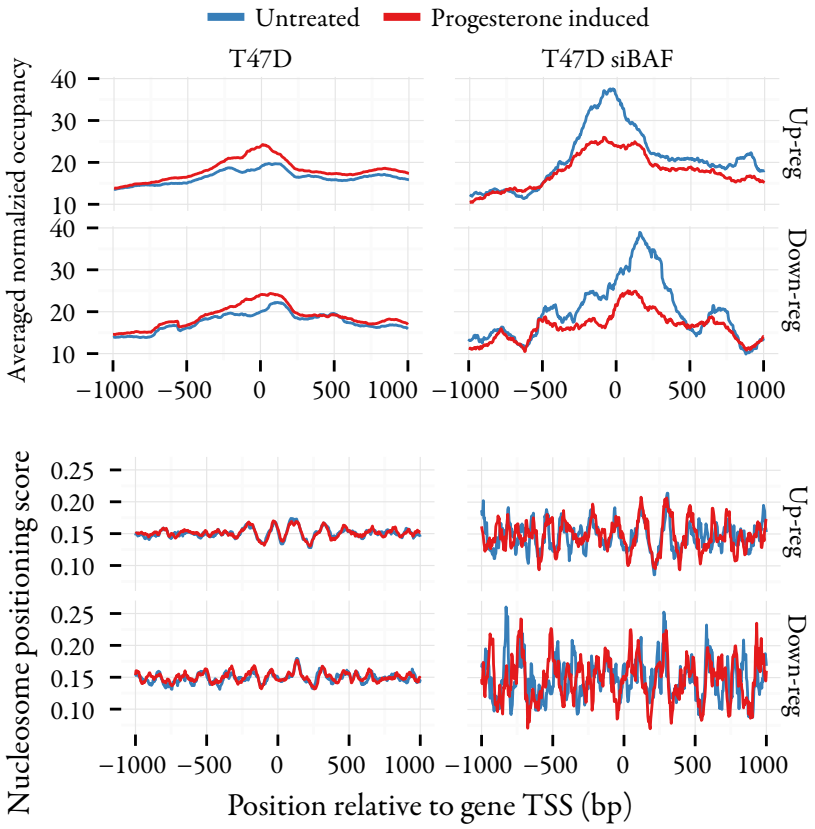


Figure S8: Nucleosome occupancy and positioning at progesterone-regulated genes. Up-regulated have higher expression after progesterone, while down-regulated have lower expression.

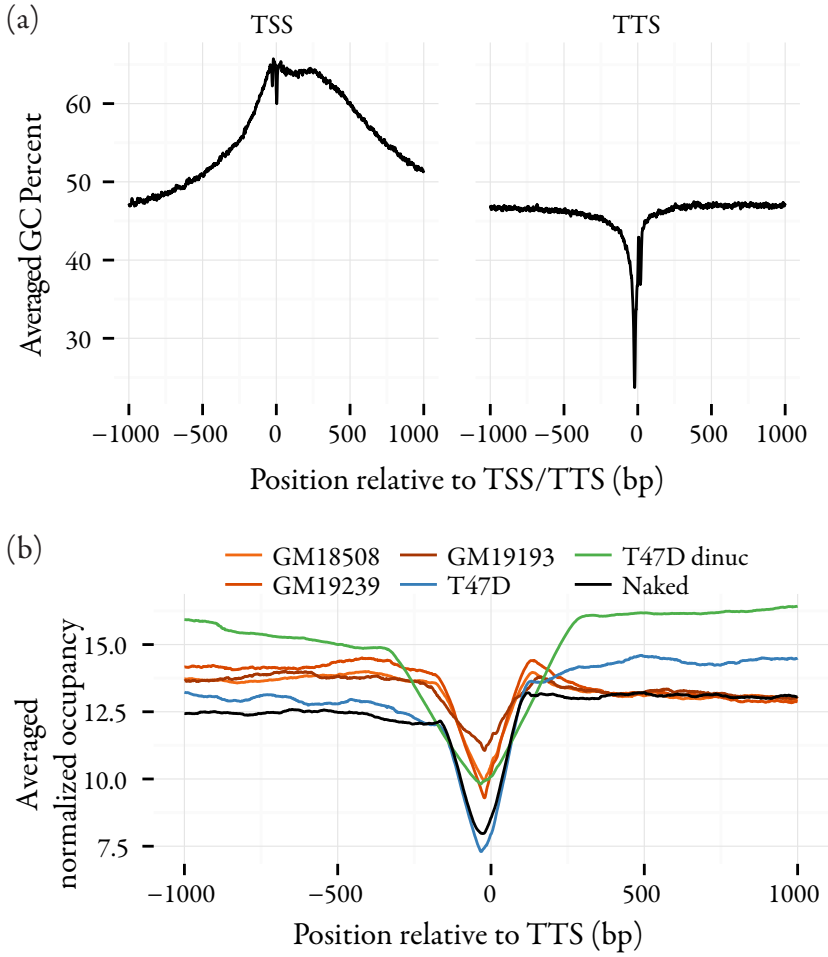


Figure S9: (a) GC Percent at the TSS and TTS of 14,561 genes. (b) Nucleosome occupancy at the TTS.

Term	Count	%	P-value
acetylation	155	26.5	4.8e-17
phosphoprotein	311	53.2	6.9e-15
ribonucleoprotein complex	45	7.7	2.5e-11
membrane-enclosed lumen	98	16.8	1.5e-10
intracellular organelle lumen	94	16.1	4.2e-10
organelle lumen	95	16.2	6.3e-10
organelle envelope	46	7.9	3.3e-9
envelope	46	7.9	2.9e-9
mitochondrian	63	10.8	3.2e-8
nucleotide binding	110	18.8	4.8e-8

Table S1: Top 10 Gene Ontology terms returned for 585 of the 610 super-occupied genes, as listed in the "Functional Annotation Chart" of DAVID [239]. The largest enriched sets of genes include those that are post-translationally modified with acetylations or phosporylations.

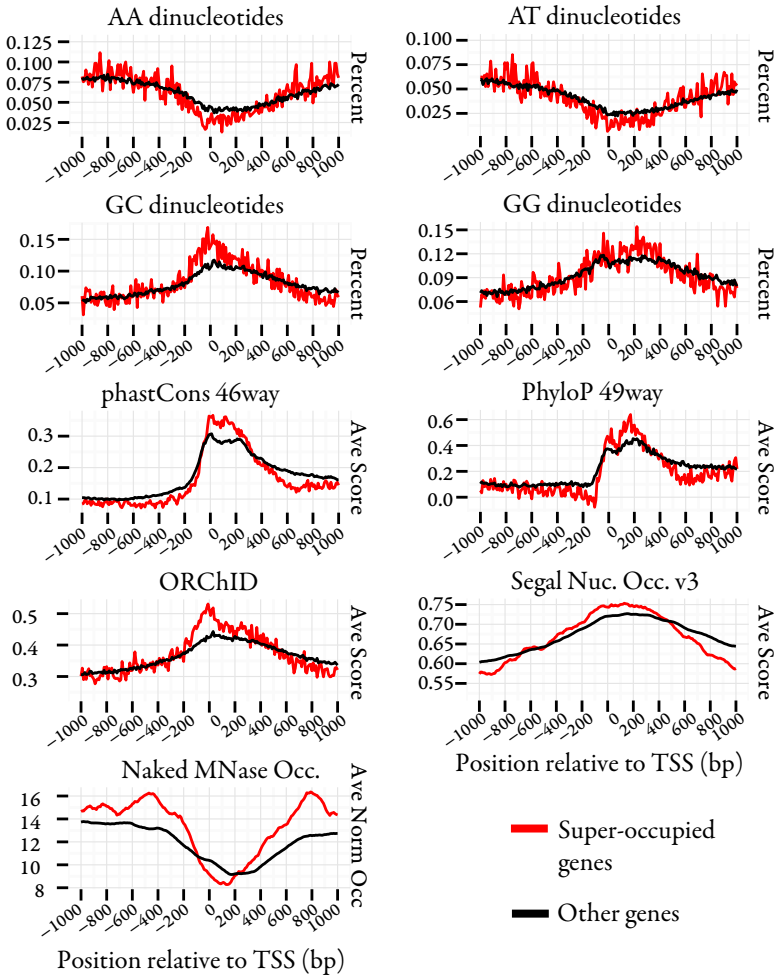


Figure S10: Different scores in the NFR. phastCons [241] and PhyloP [217] are both scores evolutionary conservation. ORChID predicts cleavage potential by hydroxyl-radicals [237]. "Segal Nuc. Occ. v3" refers to the nucleosome occupancy predictions from Kaplan, *et. al* (2008) [240].

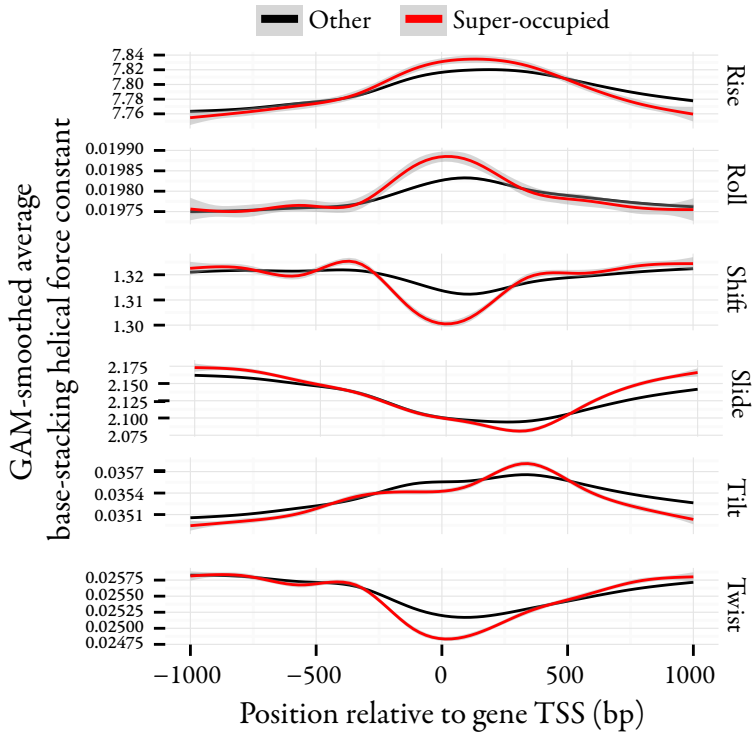


Figure S11: DNA deformability helical force constants, averaged in the regions surrounding gene TSSs and smoothed with a generalized additive model (R function stat_smooth).

	Mononuc T0	Mononuc T60	Dinuc T0	siBAF T0	siBAF T30	MNase Naked
Sequenced reads	747,392,602	787,203,218	1,281,455,389	372,865,704	371,596,738	344,800,485
Mapped reads	702,574,684	738,712,793	1,162,186,467	324,373,101	324,065,072	344,800,140
Paired frags	342,924,455	359,870,139	522,147,492	145,450,043	144,605,046	N/A
HDR region	7,947,252	9,289,283	22,330,494	2,212,569	1,962,595	9,223,494
MAPQ < 20	35,497,698	37,117,021	58,179,801	7,166,642	6,850,902	63,560,410
Bad seq. qual.	1,782,018	2,007,355	3,528,363	148,022	152,122	19,631,672
Duplicates	14,579,976	12,774,459	14,924,957	8,605,529	4,753,007	17,429,443
Reads/frags used	287,123,513	304,188,197	437,232,211	127,901,957	131,378,127	244,546,409

Table S2: Sequencing counts for nucleosome positioning data. HDR regions refer to "highly duplicated regions" from (Pickrell, *et al* 2011) [226]. Reads with MAPQ ≥ 20 according to BWA are typically mapping non-uniquely in the reference genome. Sequences deemed to have bad quality were those with 40% or more of the bases having an Illumina Phred Quality (version 1.8) of '#', which generally means the base-calling is unknown. "Duplicates" refer to the redundant set of fragments occupying the same positions. For example, if there are five fragments with the exact same start and end coordinate, one fragment is retained for analysis, while four would be discarded. A fragment may be filtered for multiple reasons, therefore the final dataset ("reads/frags used") has more fragments than the subtracting the sum of the HDR, low-MAPQ, duplicates, and bad quality reads from the number of paired fragments.

Resultats II:

Histone H1 isoform content before and after progestins

Andy Pohl, Roni H.G. Wright, Ana Silvina Nacht, Guillermo Vicent, Jofre Font-Mateu, Daniel Soronellas, François Le Dily, and Miguel Beato

(manuscript)

Abstract

Seven somatic variants of histone H1 exist in humans: H1.0-H1.5 as well as H1x. Using T47D cancer cell lines expressing different HA-tagged H1 isoforms, we examine how five H1 isoforms are distributed in chromatin, both with and without a progesterone stimulus causing the activation and repression of 4,000+ genes and widespread chromatin remodeling. The isoforms share a well-conserved globular domain, but vary significantly in their C-terminal ends, which has an impact on their affinity to bind to chromatin. While we have found a lot of redundancy among the variants, and binding preferences at the level of large-sized regions that only shift slightly after progesterone, we find more specificity at smaller-sized regions like that of gene promoters or even at the mononucleosome level. In addition, we have found nucleosome-sized sites highly-specific to each isoform and comparing their loci reveals additional patterning that also may change after progesterone induction.

Introduction

The field of epigenetics has seen incredible advances in understanding the role of chromatin in gene regulation. In particular, the heavily modified tails of core histones H3 and H4 act as switches and docking sites for factors that make nucleosomes accessible or inaccessible to transcription factors or other binding proteins. These interactions are well studied and are present in all eukaryotes. Not only are the core histones post-translationally modified, but multiple core histone variants exist: H3.3, H2A.Z, macroH2A, etc. that replace the canonical histone variant at various times, also in the context of transcriptional regulation. The linker histone H1, present mainly in higher eukaryotes, binds to the exterior of the nucleosome core particle to form the recognized nucleosome. Displacement of histone H1 is required for chromatin remodeling, as well as for the binding of polymerases at the initiation of transcription. In this way, H1 has traditionally been viewed as a repressor of gene expression with a limited role other than preserving chromatin compactness. Given the seemingly modest role H1 plays in chromatin, an interesting question arises as to why seven variants of the histone exist in somatic human cells. The variants H1.0, H1.1, H1.2, H1.3, H1.4, H1.5, and H1x are conserved particularly in their globular domain, but vary greatly in their C-terminal and N-terminal domains. Previously shown in vitro characterizations of each variant's affinity for chromatin among other properties of the proteins has concluded that the variants with longer C-terminal tails are more tightly bound to chromatin [50].

Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) is now a standard technique for detecting, genome-wide, where specific proteins are bound to DNA. Although fast sequencing makes the technique possible, the real power of ChIP-seq comes from the antibody. Antibodies used in ChIP experiments have been raised against a multitude of chromosomal proteins, and are even specific to post-translational modifications. Sadly, not all antibodies have equal specificity. To overcome the limited availability of ChIP-grade H1 isoform-specific antibodies, we previously constructed T47D cell lines expressing five of the seven somatic H1

variants, each with a human influenza hemagglutinin (HA) epitope tagged to the C-terminal end of the protein [57]. Although expressed under a strong CMV promoter, the infected cell line clones were carefully selected in such a way that the HA-tagged H1s express a similar level of protein as their endogenous counterparts.

We use the T47D cell line due to its expression of a functioning progesterone receptor, and inducing cells with progesterone (or a synthetic progestin) will cause widespread changes in chromatin. Briefly, the activation model has two stages. First, H3 is phosphorylated at serine 10 leading to the eviction of an HP1-containing complex and the recruitment of an ASC2/MLL complex, leading to H3K4me and the recruitment of a NURF complex [223], in turn leading to the recruitment of PARP1 and Cdk2/CyclinA, which phosphorylates histone H1, promoting its displacement [242]. In the second stage (lasting 5-10 min), the H1-depleted nucleosomes allow PR-BAF complexes to bind and mediate ATP-dependent displacement of H2A/H2B, and the subsequent binding of NF1 [223]. Bound NF1 stabilizes the open conformation of the chromatin, resulting in activation of the promoter. The open conformation of chromatin peaks after 30 minutes of progesterone treatment. Other recent studies have examined genome-wide H1 isoform binding [52,243,244] but ours is the first to look at changes in H1 deposition in response to a stimulus: in this case, regulation of gene expression and chromatin remodeling caused by the steroid hormone progesterone.

Materials and Methods

Cells

T47D-HA breast cancer cells carrying HA:H1 variants under the control of the CMV promoter were established in previous work [245]. For the MNase-seqe, RNA-seq, DNase-seq, and non-HA ChIP-seq experiments, we used T47D cells expressing a single copy of the MMTV promoter, and are also known as T47D-MTVL cells. These cells were derived from a single clone named "3/17".

We performed some diagnostics to ensure proper functionality of the HA:H1 cell lines. We checked their morphology, HA:H1 protein expression, and function of the MMTV promoter.

Immunofluorescence

T47D cells were grown on glass 12mm round coverslips. 24 hours prior to progesterone induction growth medium was replaced with RPMI minus FCS. R5020 (10mM) was then added for the desired length of time and the coverslips fixed using 4% paraformaldehyde in PBS for 15 min and permeabilized with PBS 0.2% Triton X-100 at room temperature. Coverslips were then blocked with 5% skim-milk for 1 h at room temperature and incubated for 2 hours with primary antibodies diluted in PBS 5% skim-milk at 1/500 (mouse anti-HA antibody sigma cat: 9658). Following three washes with PBS Tween 20 0.05%, coverslips were incubated with secondary antibodies (AlexaFluor 488 anti-mouse, Invitrogen-Molecular Probes) for 1 hour at room temperature. After an additional three washes with PBS Tween 0.05% and DNA staining with DAPI, samples were mounted with mowiol and the images were acquire with a Leica TCS SP5 CFS confocal microscope.

Chromatin immunoprecipitation

Following R5020 treatment, medium was replaced with medium lacking R5020 (also serum-free and without phenol red), and proteins were crosslinked to DNA by adding the crosslinking solution (1% formaldehyde), directly to the culture medium and incubating for 10 minutes at 37°C. Then, the crosslinking reaction was stopped by adding Glycine at a final concentration of 0.1M and incubating the plates for 5 minutes at room temperature. The medium was removed and cells were washed twice with cold 1x PBS containing protease and phosphatase inhibitors (1 mM Phenylmethylsulfonylfluoride (PMSF), 1 μg/ml aprotinin, 1μg/ml pepstatin A, 1 μM Sodium Ortovanadate, 20mM β-Glycerolphosphate and 1X protease inhibitory cocktail (PIC) from Roche). Cells were scraped in the presence of PBS-containing inhibitors and centrifuged for 5 minutes at 4000 rpm at 4°C. Cell pellets

were then resuspended in 2.5 ml of cell lysis buffer containing inhibitors and incubated for 10 minutes on ice. Followed this first lysis, cells were pelleted for 5 minutes at 4000rpm at 4ºC then were resuspended in 0.5 ml of a nuclei lysis buffer solution. The lysate was sonicated on ice to yield DNA fragments between 150 and 200 bp. After sonicating, the material was cntrifuged for 10 minutes at 13000 rpm at 4ºC and cell debris was removed from the supernatant. An aliquot of this chromatin was treated overnight with Proteinase K at 65ºC. DNA was then recovered by phenol-chloroform extraction followed by its precipitation with 10% of Sodium acetate and 1.5 volumes of pure Ethanol. DNA was quantified using a Nanodrop spectrophotometer, and the size of sheared DNA was visualized on a 1.2% agarose gel. Chromatin inmunoprecipitation was performed using 30 µg of chromatin per sample and diluted 1:10 in ChIP IP buffer containing protease and phosphatase inhibitors at the same concentration described above. For input control, 50 µl of this diluted chromatin were recovered before adding the anti-HA antibody (Abcam cat. no. ab9110). In order to perform the Inmunoprecipitation, 5µg of the antibody were added to the diluted chromatin and incubated overnight at 4ºC on a rotator. The day after the incubation with the antibody, 42 µl of One Day, Ab binding protein A agarose (Diagenode cat. no. kch-503-008) were added to each reaction after blocking it with 0,5% of Serum Bovine Albumin (BSA) at 4ºC while rotating on a wheel for 15 min. The antibody-containing chromatin suspension was incubated with the protein agarose beads for 3 hours at 4ºC with rotation. After incubating, the beads were pelleted by gentle centrifugation (2 min. at 3000rpm, at 4ºC) and supernatant containing unbound unspecific DNA was discarded. The agarose with bound antibody/protein/DNA complexes was washed three times for 5 minutes at 4ºC with rotation with 1X ChIP Buffer Diagenode (cat. no. kch-501-700) followed by two washes with Tris-EDTA buffer 1X (TE1X). The DNA was eluted by incubating twice the washed agarose with elution buffer for 15 min. at room temperature with rotation. Crosslinking was reversed by incubating samples overnight with 0.2M of NaCl at 65ºC. Proteins were then digested by incubation for 1 hour at 45ºC with Proteinase K and DNA was recovered by phenol-chloroform extraction. DNA was precipitated with 200 mM of NaCl,

1.5 volumes of pure ethanol and 0.1% of glycogen, washed once in 70% of ethanol and DNA pellet dissolved finally in 25 µl of DNase-free water.

Buffers

- Crosslinking solution: 50 mM HEPES pH 8.0; 0.1M NaCl; 1 mM EDTA pH 8.0; 0.5 mM EGTA pH 8.0.
- Cell Lysis Buffer: 5 mM PIPES pH 8.0; 85 mM KCl; 0.5% NP40.
- Nuclei Lysis Buffer: 1% SDS; 10 nM EDTA pH 8.0; 50 mM Tris-HCl pH 8.0.
- 1X ChIP Buffer Diagenode: 5X ChIP Buffer Diagenode diluted 1:5 in water.
- Elution Buffer: 1% SDS; 0.1M NaHCO3.

Antibodies for non-HA ChIPs

CTCF (Millipore cat. no. 07729), H3K27me3 (Active Motif cat. no. 39155), H3K14ac (Millipore cat. no. 07-353), endogenous H1.2 (Abcam ab4086), HP1gamma/CBX3 (Millipore cat. no. MAB3450), PR (Santa Cruz cat. no. H190), p300 (Santa Cruz cat. no. sc584/5), RAD21 (Abcam cat. no. ab992), H2A.Z (Abcam cat. no. ab4174), FOXA1 (Abcam cat. no. ab5089), H3K-27ac (Abcam cat. no. ab4724).

RNA-seq and DNase-seq

RNA-sequencing was performed by collecting purifying total RNA from untreated cells and cells treated with progesterone for 6 hours. Ribosomal RNA was depleted using the Ribominus kit (Invitrogen), then remaining RNA was used to construct Illumina paired-end sequencing libraries, then sequenced 2 x 50 nucleotides. Sets of genes for all of the analysis were selected based on the RNA-seq.

DNase-seq was performed using the protocol from Song, *et al.* (2010) [166], with modifications of the protocol to accommodate T47D cells.

Mass Spectrometry

T47D cells were grown in usual way on six 150 mm plates, to around 70% confluence. To extract histone H1 proteins, we used an Active Motif Histone Purification Kit (cat. 40025), following the special instructions on extracting histone H1 separately from the core histones. Samples were digested with LysC/Trypsin. 2 μg of the sample was analyzed by LCMSMS using a SHORT_CID method in the nanoLC LTQ Orbitrap Velos XL (Thermo Scientific). To avoid carry over, BSA runs were added between samples. BSA controls were included both in the digestion and LC-MS/MS analysis for quality control. Resulting peptides were searched against the SwissProt human database [246], using an internal version of the search algorithm Mascot (http://www.matrixscience.com/). Peptides were filtered based using an FDR > 1%. Proteome Discoverer v1.4 (Thermo Scientific) was used to assign peptides to individual proteins. Proteome Discoverer gives an approximate estimation of protein amount with the parameter "Area" which is the average peak area of the 3 top peptides for a given protein. Using resulting approximate quantifications averaging two replicates, relative proportions could then be inferred.

Genes and transcription start sites

To define our list of TSS regions, we compiled a list of protein-coding genes from GENCODE v19 (20,318 total). Among this list of genes are a number of genes with little or no expression (RPKM < 0.002) seen in RNA-sequencing experiments we performed in the same conditions. We removed these genes, as well as genes with very complex loci, in order to have a set of TSSs (14,561 total) we could perhaps analyze more efficiently.

Progesterone-regulated genes

In addition to this main set of genes, we also categorized the genes with differential mRNA expression before and after 6 hours of progesterone treatment. 1,046 genes were up-regulated, or had higher expression (a fold-change of > 1.5) after progesterone than before. Likewise, 587 genes were down-regulated

(negative fold change > 1.5) by progesterone. Finally, 4,324 genes that were expressed at a consistent level before and after progesterone (-1.2 < FC < 1.2) were considered to be non-regulated. We had previously reported around 2,000 genes both up (FC > 1.5) and down-regulated (FC < -1.2) by progesterone [146] as measured by microarrays. We used the RNA-seq experiments to cull this list, selecting genes consistently up, down, or non-regulated by both techniques, as well as restricted our defining thresholds to better discriminate the three classes of genes.

shH1.2 knockdown genes

Genes affected by the doxycycline-activated knocking down of histone H1 variant H1.2, were reported previously in the T47D cell line [245]. In this case, only 9 genes were down regulated by the knockdown of H1.2, and 54 were up regulated. In this case, the regulation is not known to be influenced by progesterone, as that was not used in the experiment.

Sequence processing

Sequenced H1 samples were subjected to a similar pipeline as the MNase-sequencing pipeline from the previous chapter. Briefly, they were aligned to the human genome reference sequence GRCh37/hg19, filtered for bad sequence quality, multiple mapping, duplicated fragments, mapping to highly-duplicated regions in the reference [226], and paired-end reads having problems with the mate read were also discarded. Table S1 lists the results of filtering the sequence. In general, we use the term "fragments" to distinguish properly-mapped paired-end reads from mapped single-end reads.

Genome-wide H1 occupancy

After compiling raw fragment depths, we calculated a normalized fragment depth based on the input sample's fragment depth and the mean input depth in a local 20 kb window. We then subtracted the normalized input's fragment depth from each normalized sample's. The input-subtracted normalized fragment depth is the data used in much of the analysis focused on aggregated depth profiles at genes and positioned nucleosomes.

Relative Ratios

The relative ratio of a sample *j* at each base *i* of the genome was calculated as the normalized depth of a sample divided by the sum of normalized depths for *m* number of samples:

$$RR_j(i) = \frac{NormD_j(i)}{\sum_{k=1}^{m} NormD_k(i)}$$

Mappability is left out of the equation in order to preserve the property that the relative ratios of all samples at a given base will sum to one.

Fragment counts

Prior to counting fragments, sample datasets with higher numbers of total fragmens were downsampled to the sample with the lowest number of fragments (H1.3 T30, 1.02×10^8 total fragments). Counts per 100 kb region of the genome were made for each sample using a program called "bambed" (unpublished). Fragments lying across the boundary of each 100 kb bin were included in the count of the bin containing more than 50% of the fragment. Fragments equally contained in two bins were counted in the upstream bin.

Entropy and information content

Rough-scale competitive chromosomal deposition of H1 isoforms was calculated using the downsample-normalized fragment counts for each isoform in 100kb non-overlapping windows, genome-wide. In each interval, the isoform with the highest normalized read volume was declared the representative isoform. As a measure of isoform purity over an interval, the information content was calculated by first converting the summed normalized read depths (volumes) into relative proportions (p_k) then using the following formula:

$$IC = \log_2 5 - \sum_{k=1}^{m} p_k \log_2 p_k$$

The value of the information content ranges from 0 to $\log_2 5 \approx 2.32$. A low information content indicates more equal proportions, while higher values

mean a disproportionate amount of one or perhaps two of the subtypes is present.

Selecting well-positioned nucleosomes

Beginning with the calculated genome-wide nucleosome positioning scores taken from another project (see Results I chapter's methods for a description of the score), for a particular sample, we first gathered bases harboring local maxima in perfectly mappable regions at least 2kb or more in size, including 200bp buffers on either end of the region. In a greedy manner, we then selected the highest maxima no fewer than 150 bp from another local maxima (using "bwtool find" with parameters –maxima and –min-sep=150), resulting in in a single base that we accept to be the best-positioned dyad locally. Because the score itself is invariant to nucleosome occupancy, and low-occupancy can lead to artificially-high scores, we also filter out nucleosomes with occupancy lower than that sample's mean occupancy. Finally, we chose nucleosomes from samples having normalized occupancy 1.5 fold or more higher than that of the normalized occupancy of MNase-treated naked DNA. Well-positioned dinucleosome centers were found in the same manner, although we used a separation parameter of 300 instead of 150.

Results

H1 variant deposition at the level of chromosomes

The H1 isoforms have an uneven distribution within the nucleus of the cell. In order to analyze this distribution in T47D cells, we performed immunostaining experiments using an HA antibody in each of the cell lines expressing individual H1 somatic variants(Figure 1). The stains show stark differences in localization, with H1.4/H1.5 localized in the nuclear periphery, and H1.0/H1.2/H1.3 more widespread. Like most terminally-differentiated cells, H1.1 is virtually absent in T47D cells.

T47D cells express the H1 variants at different levels. The immunoprecipitated ChIP DNA is sequenced to around the same number of frag-

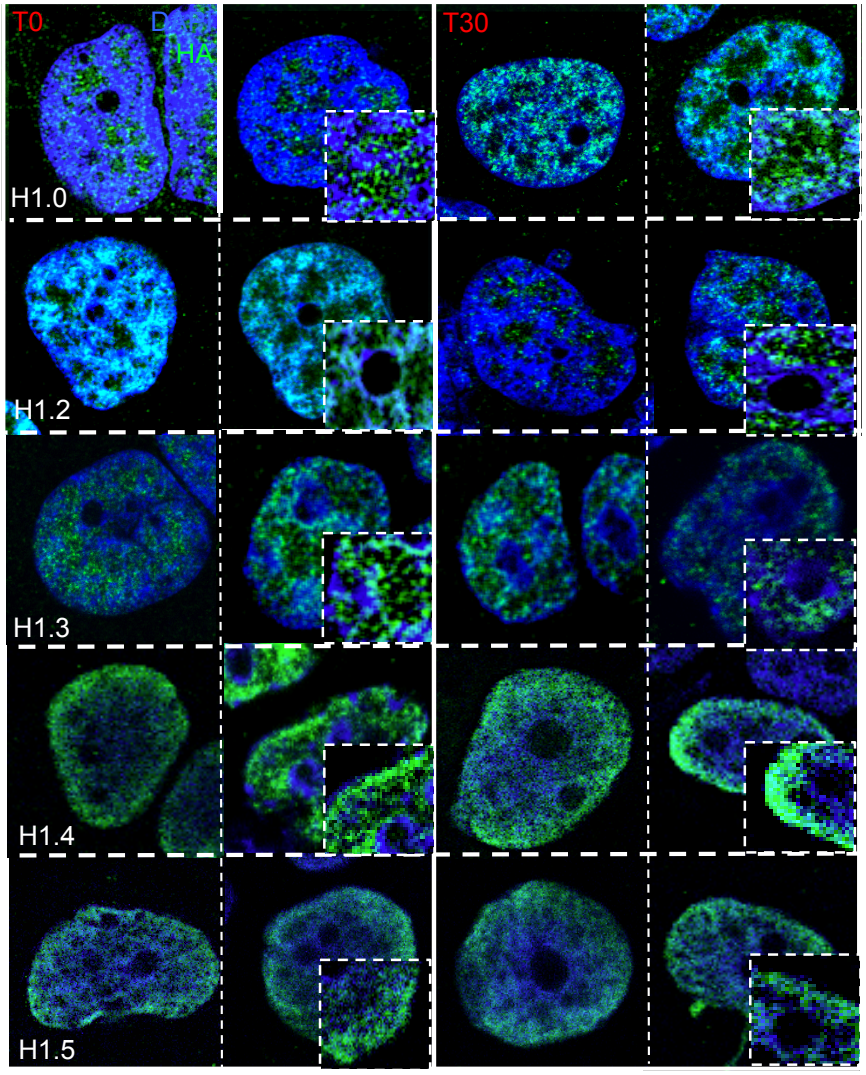


Figure 1: Immunostaining HA:H1 variant cell lines using the a primary antibody against the HA peptide, with (right panels) and without (left panels) a progesterone treatment for 30 minutes.

ments in each sample, giving the impression that the H1 isoform proteins are present in roughly equal proportions in the nucleus. Although the actual rough proportions were reported previously [57], we performed mass spectrometry experiments to get a better estimate on endogenous levels of the protein. In T47D cells, H1.5 is by far the most present protein, comprising around 45-49% of the H1 content (Table 1). At the low end, H1.3 accounts

Isoform	Experiment 1		Experiment 2		Average
H1.2	1,85E+10	18%	2,20E+10	20%	19%
H1.3	6,08E+09	6%	4,69E+09	4%	5%
H1.4	1,84E+10	18%	1,10E+10	10%	14%
H1.5	4,79E+10	46%	5,44E+10	49%	47.5%
H1.0	1,40E+10	13%	1,80E+10	16%	14.5%

Table 1: Peptide counts and percentages from mass spectrometry experiments (technical replicates).

for around 5% of total H1.

As seen by immunostaining, the effect of progesterone on the localization of individual variants is weak: in general the isoforms localize to the same regions of the nucleus after a progesterone treatment. Using a ChIP-grade antibody against the HA tag, we performed chromatin-immunoprecipitations of HA:H1 in T47D cells that were also treated or not treated with progesterone for 30 minutes. The DNA recovered from these ChIPs was sequenced, and those sequences were aligned to the reference human genome (GRCh37/hg19). As a comparison of the general dispersion of fragments, we downsampled the mapped datasets to be equal in size to the dataset with the fewest fragments (H1.3 T30, see Table S1), then compared the genome coverage, and the mean and standard deviation of the fragment depth (Table 2). Even with slight variation to the fragment sizes (Figure S4), it is striking to see that while H1.0 and H1.4 cover about the same amount of the genome before and after progesterone, H1.2 and H1.3 cover about 50 megabases more after hormone, and H1.5 covers about 50 megabases less.

Genome-wide profiles of the raw read depth for each sample were generated, normalized by scaling to the average input depth, and then subtracting the raw input depth, and multiplying by a mappability value [228]. These signal data were used more for analyses at specific loci, but at the genome-wide level, pairwise Pearson correlations between samples were calculated using 1.2 gigabases of regions of perfect mappability 2 kb or more in length, with 200 bases removed from each end to avoid border effects. A summary of the reduction of the size of the perfectly-mappable genome at varying fragment lengths is shown in Supplementary Figure S2. The cor-

Isoform	Proges-terone	Genomic coverage (gigabases)		Mean frag-ment depth	Std. dev. frag depth
H1.0	T0	2.647	—	6.887	4.797
H1.0	T30	2.654	—	7.026	4.980
H1.2	T0	2.578	↓	7.022	5.271
H1.2	T30	2.628	↓	7.077	5.187
H1.3	T0	2.582	↓	6.945	4.977
H1.3	T30	2.632	↓	7.021	4.798
H1.4	T0	2.629	—	6.673	4.694
H1.4	T30	2.622	—	7.015	5.147
H1.5	T0	2.638	↑	6.724	4.652
H1.5	T30	2.597	↑	7.428	5.522

Table 2: Genomic coverages and mean fragment depth of each sample, downsampled to match the sample with the fewest number of fragments.

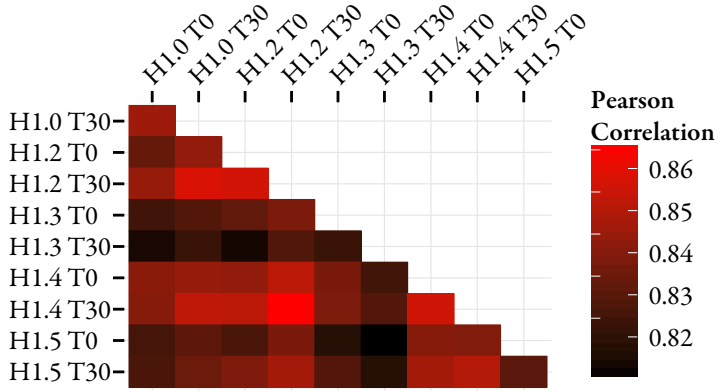
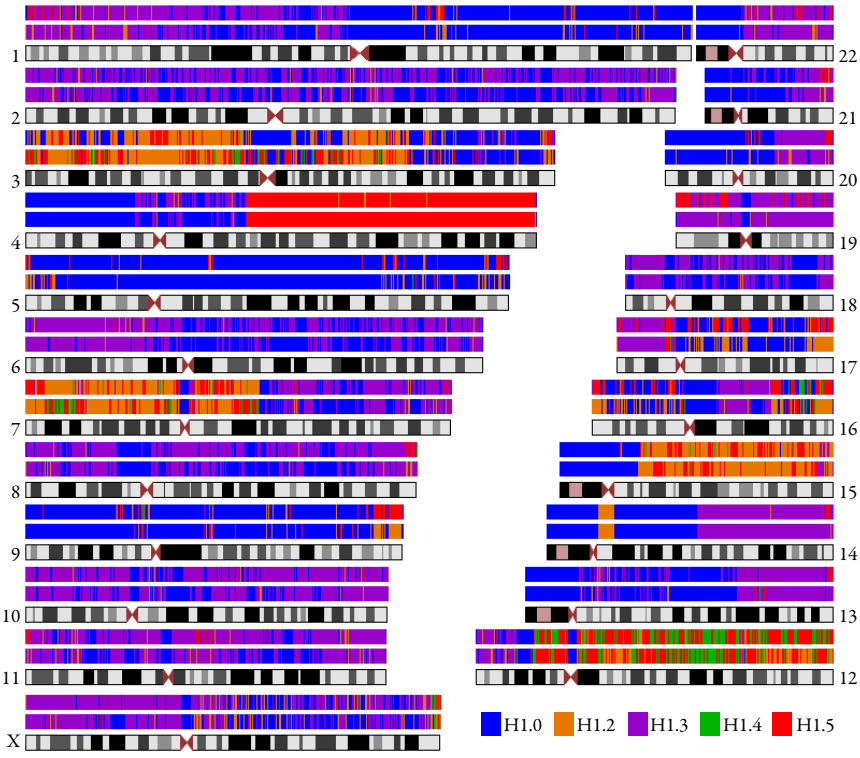


Figure 2: Pairwise Pearson correlation coefficients calculated between samples genome-wide, in the 2 kb perfectly-mappable regions. The regions show H1.3 at both timepoints having the least correlation to the other samples.

(a)



(b)

Isoform	T0	T30
H1.0	474	467
H1.2	15	7
H1.3	319	233
H1.4	92	41
H1.5	29	209

(c)

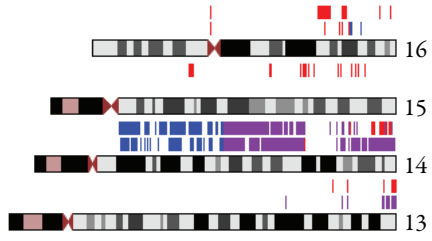


Figure 3: **(a)** Genome-wide view of H1 binding. Above each chromosome are two lines: the line immediately above a chromosome shows binding at without progesterone (T0). The line above the T0 line shows the binding after 30 minutes of progesterone (T30). In 100 kb intervals, the variant with the most reads was plotted as the representative H1 in that interval. Although rough binding preferences at the 100kb scale are sometimes clustered into long segments, and are largely conserved before and after hormone, the underlying percentages of each variant are very similar. After calculating the information content for each 100 kb bin, the resulting winners were reduced to the set listed in the inset table **(b)**, after setting the minimum information content to be greater than 0.01. **(c)** After setting this threshold, much of the diagram in (a) becomes blank, but a few chromosomes, particularly chromosome 14, still display strong isoform-specific preferences.

relations between samples are plotted as a heatmap in Figure 2. By and large, the samples correlate highly to each other, however differences can still be seen, and H1.3 presents the most unique profile, particularly after treatment of progesterone.

We segmented the genome into bins 100 kb in size and for each bin calculated the sum of fragments from each sample, after downsampling each dataset to equal the dataset with the lowest number of fragments (in our case 101.6 million, see Table S1). For each bin, the #1 variant (that with the highest number of fragments), was plotted in Figure 2. The "winner" bins tend to cluster together in consecutive 100 kb segments giving rise to very large regions, in some cases 100 megabases, e.g. H1.5 on chromosome 4. Visualizing the deposition of H1 isoforms in this manner is also a way to highlight changes in deposition after hormone. Although nearly two-thirds of the 100 kb regions have the same winner after progesterone, various exchanges have been categorized in Table S3. The isoform notable in this case is H1.2. For the other isoforms, the highest transition count appears on the diagonal, i.e. regions associated with the subtype before progesterone, continue to have that association after progesterone. In the case of H1.2 though, many of the associated regions become H1.5-associated after progesterone. Although initially compelling, the underlying differences in isoform fragment counts per 100 kb bin are quite subtle. In order to quantify the differences among the variants as a single number per 100 kb bin, we calculated the Shannon entropy and corresponding information content of their proportions. The information content of a 100 kb bin with perfectly equal proportions amongst the isoforms will have a value of zero, and in our case in the T0 sample only 929 (3%) of the 30,378 100 kb bins had an information content > 0.01. Likewise, in the T30 sample, only 957 bins (3.2%) were > 0.01. The inset table in Figure 3 lists these bins per isoform. Interestingly, the bins with elevated information content were largely restricted to just several chromosomes, including chromosome 14, as seen in Figure 3c.

H1 variant deposition surrounding genes

From RNA-sequencing experiments taken with and without progesterone

treatment, we mapped paired-end reads to 14,561 genes with a minimum RPKM (reads per kilobase per million) [193]. We sorted these genes by RPKM then divided them into five groups of equal size. Aggregated profiles of the 2 kb region surrounding each gene's transcription start site (TSS) and transcription termination site (TTS) were calculated for each H1 sample's input-subtracted normalized read depths (Figure 4). In the case of all samples, progressively lower gene expression led to progressively higher levels of histone H1. All samples also demonstrate a region of roughly 500 bases near the transcription start site highly depleted of histone H1 at all levels of expression apart from the group of genes with the lowest gene expression, which actually showed slight enrichments in these regions specifically. To examine the differences in isoforms specifically, we also plotted the subtraction of the T0 plot from the T30 plot. The subtracted profiles of H1.0, H1.2, and H1.4 are rather flat, and it is quite clear that binding overall level of those variants do not change after treatment of progesterone. By contrast, after progesterone, H1.3 has reduced binding and H1.5 has increased binding. In all cases, differential binding is proportional to level of expression in the genes. The TTS regions show a similar effect to the TSS regions albeit a bit more subdued, but interestingly show more local depletion in the lowly-expressed genes than in the TSS. Another observation from the TSS plots is that some of the isoforms bind asymmetrically across the nucleosome free region (NFR). In particular, H1.2 T0, H1.3 T0, and H1.5 T30 have increased binding in the +1 nucleosome compared to the -1 nucleosome.

We also looked at the TSS/TTS regions of several other sets of genes. Sequencing the mRNA in untreated and progesterone-treated cells, we found 1,046 genes up-regulated (FC > 1.5) by hormone, 587 down-regulated (-FC > 1.5), and 4,324 genes with similar expression before and after hormone (FC < 1.2). Similar to the other genes, we plotted the 4 kb region surrounding the TSSs (shown in Figure 5). While the effect of progesterone is quite strong on the redistribution of H1 isoform binding, the same basic pattern is observed before and after progesterone in all three different classes of genes: H1.0, H1.2, and H1.3 all have less binding after progesterone, while H1.4 and H1.5 have more, particularly H1.5. Differences between the three classes

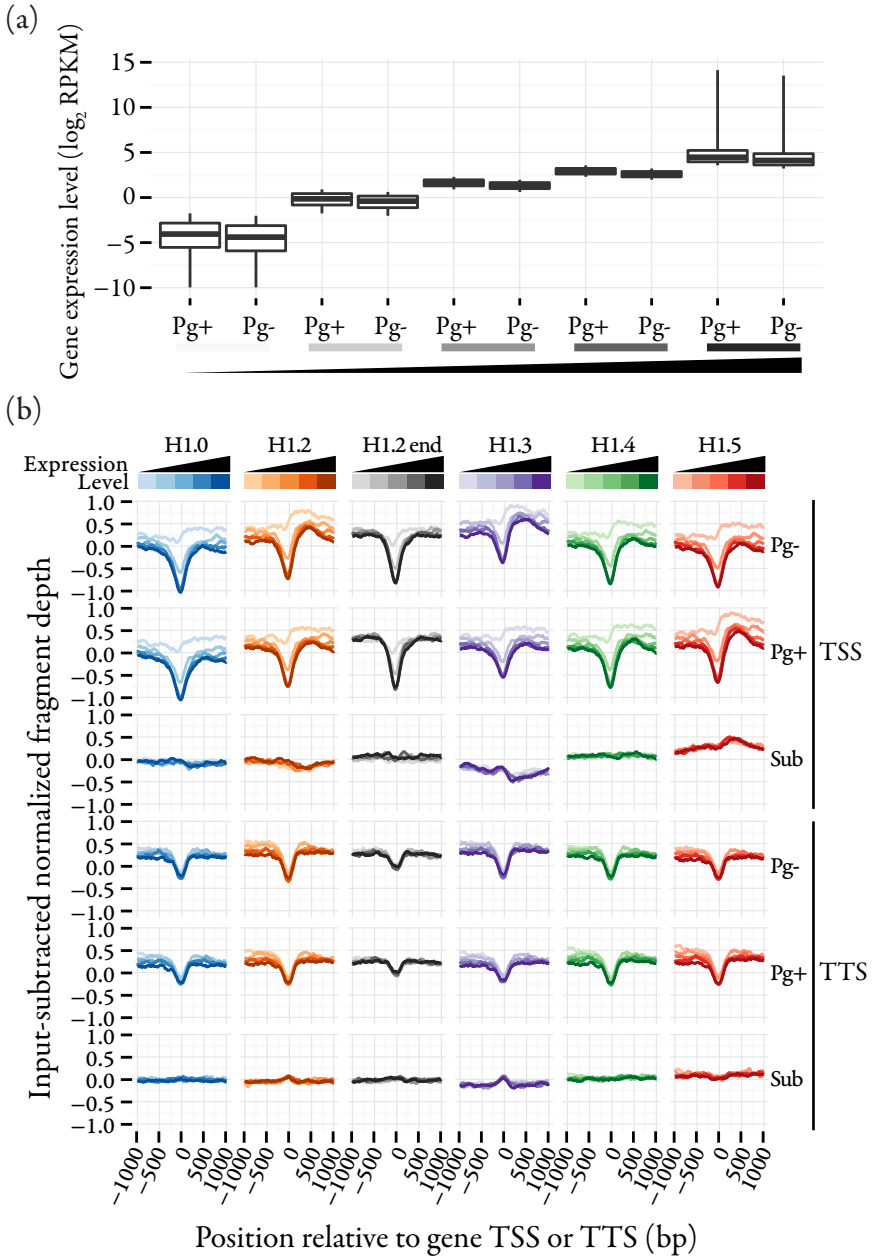


Figure 4: H1 isoforms and level of transcription. (a) The total set of genes divided into fifths, ranked by expression, separately for the RNA-seq experiments before and after progesterone induction. The gene expression distributions for the resulting quintiles are shown. (b) Normalized and input-subtracted fragment depth profiles for each H1 isoform, averaged and centered at the TSS or TTS of genes within each expression quintile, before and after progesterone. In addition, the third and sixth rows of plots show a subtraction of the before-progesterone plot from the after-progesterone plot.

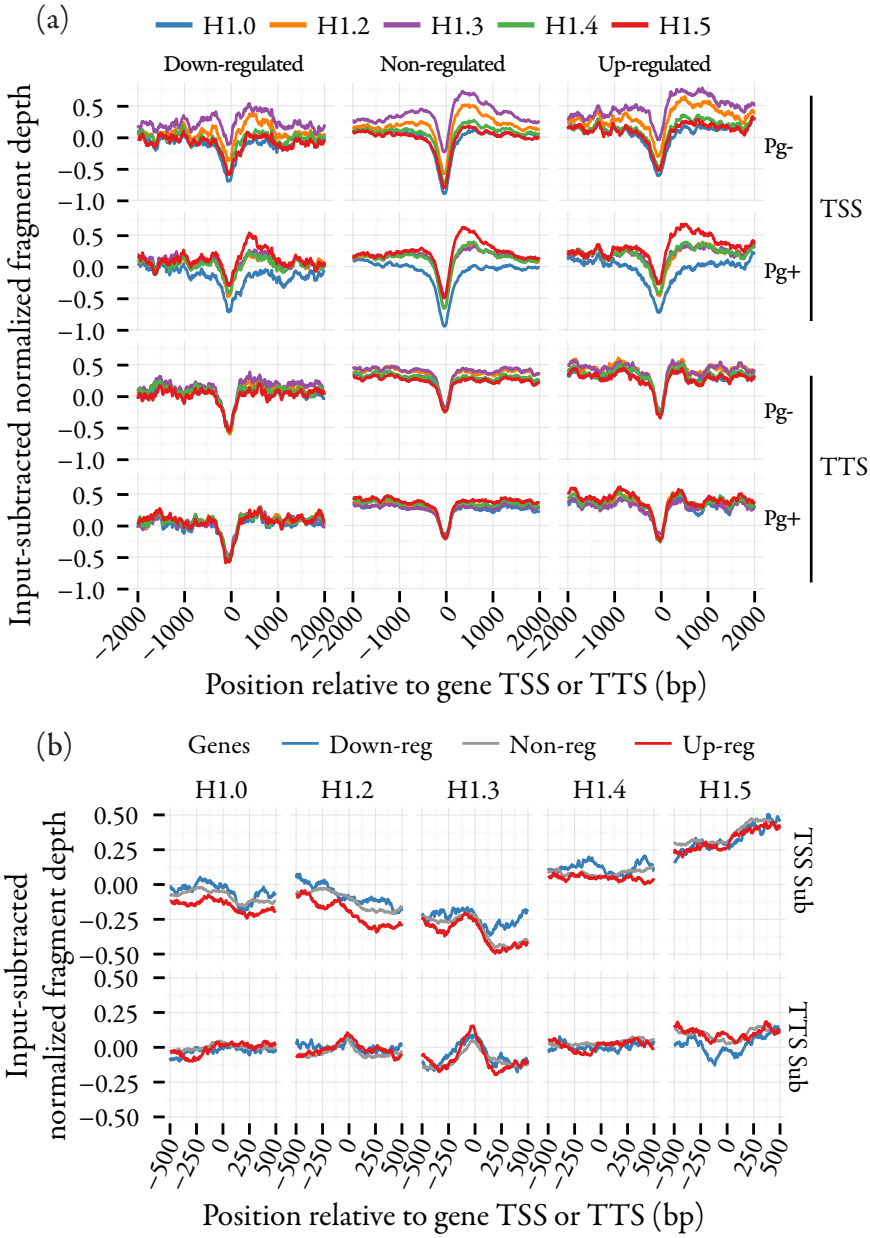


Figure 5: H1 isoforms at the TSS and TTS of genes regulated by progesterone.

isoform exchange after progesterone follows the pattern seen in the regulatory regions of the various gene groups.

To find regions of H1 isoform specificity on a finer scale (nucleosome resolution), we used HOMER [198], a flexible pipeline for finding peaks in

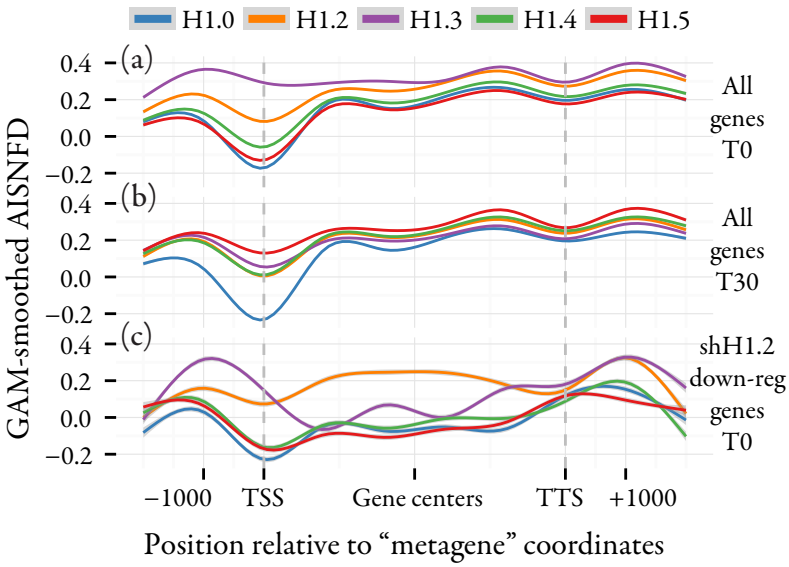


Figure 6: Metagenes profiles of (a) all genes before progesterone, (b) all genes after progesterone, and (c) shH1.2 down-regulated genes before progesterone.

of genes are perhaps highlighted best in Figure 5b. Here, subtracted profiles reveal differences between up and down-regulated is most profound in H1.2 at the TSS.

Knowing that important epigenetic signals occasionally arise in the gene body as opposed to the promoter region (e.g. H3K36me3), we performed similar aggregate analysis as in the TSS/TTS, except first scaling the entire gene body to a 5 kb region known as a "metagene". An increase in H1.5 after progesterone is seen not just surrounding the TSS/TTS, but throughout the gene body, as seen in Figure 6a and 6b, using the entire set of 14,561 genes. Microarrays were performed previously [245] using the T47D cell line, with a doxycycline-inducible shRNA targeting histone H1.2 transcripts. A list of 54 genes was collected from these microarrays where the level of expression was significantly less after knocking down H1.2. We calculated the metagene profile for the binding of H1 at these genes without progesterone, and found depletion of all variants other than H1.2 across a large portion of the metagene body (Figure 6c).

H1 variant deposition at the nucleosome level

Region/Peak description	Mono T0	Mono T60	Dinuc T0
2 Kb perfect mappability	1.19 Gb	1.19 Gb	1.09 Gb
with mean occ.	454.8 Mb	487.0 Mb	373.1 Mb
peaks ≥ 0.6, 150 bp† sep	3,265,702	2,645,324	479,283
filtering mapp./occ.	172,778	119,437	6,565
filtering naked occ.	39,432	37,481	**2,886***
peaks ≥ 0.75	7,559*	5,953*	228
change T0/T60 ≥0.4	**1,054**	**576**	N/A
exonic peaks (% overlapping starred sets)	818 (10.8%)	613 (10.2%)	692 (24.0%)

Table 3: Selection of well-positioned nucleosomes in MNase-seq samples begins by reducing the genome to regions with good mappability and where the raw fragment depth (occupancy) is at least with scoring each base in regions of good mappability. Many bases harbor localized peaks of this score, so peaks are chosen using a greedy algorithm to ensure that every dyad chosen is the best in the 150 bp span surrounding it.

From MNase-sequencing experiments we derived a set of well-positioned nucleosomes (see Table 3), both with and without progesterone induction (albeit 60 min of progesterone treatment instead of 30), and plotted the average H1 binding at these sites (Figure 7a). At the well-positioned nucleosomes that lose positioning after progesterone, there is a relative depletion of histone H1.2 prior to hormone and a relative enrichment of H1.3 after hormone. At the well-positioned nucleosomes that gain their positioning after hormone, there is no particular isoform that stands out, however the larger region is highly depleted of all H1 subtypes, more so than the nucleosomes better positioned prior to hormone. In another MNase-sequencing experiment, instead of mononucleosomal DNA, we sequenced dinucleosome fragments. We applied a similar methodology as with mononucleosomes to seek dinucleosomes that are particularly well-positioned. Despite not having a progesterone-treated dinucleosome positioning sample, we found a much different pattern of H1 binding. Well-positioned dinucleosomes are located within regions of high H1 occupancy among all isoforms, punctuated at the center of the dinucleosome (Figure 7b). At dinucleosomes, the pattern of H1

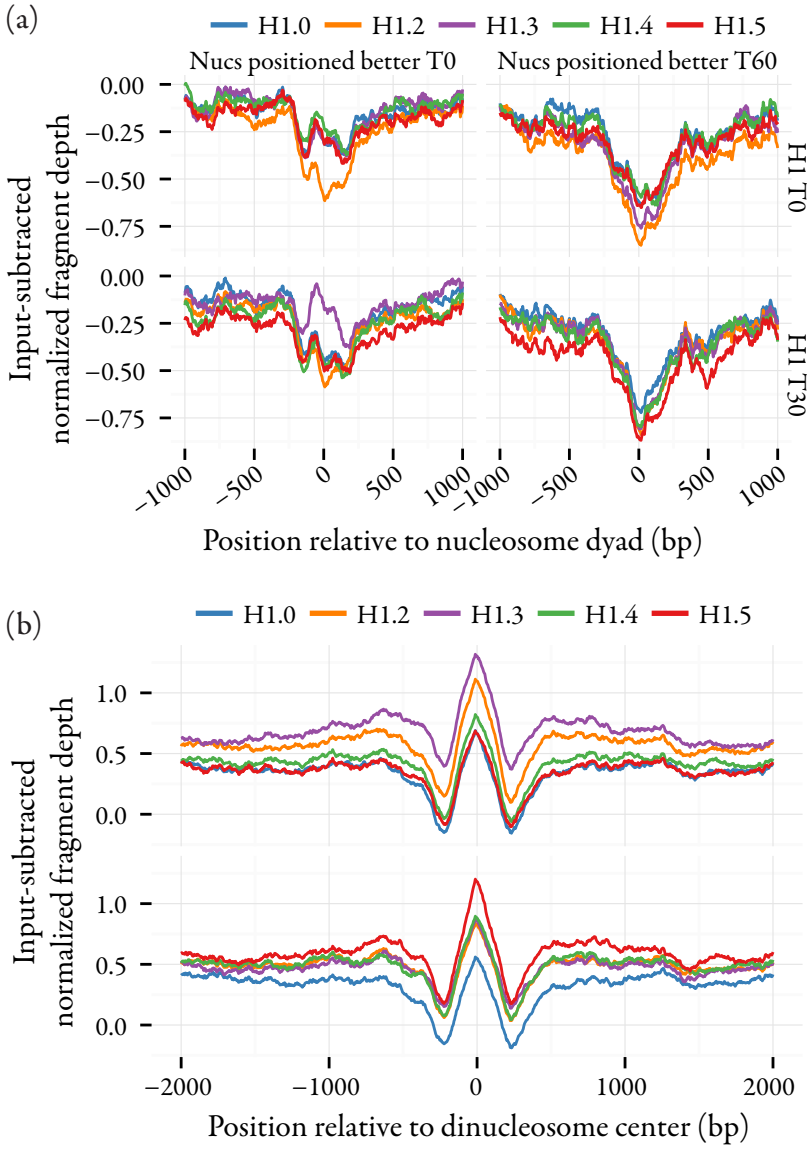


Figure 7: H1 isoform binding at well-positioned nucleosomes from Table 2. (a) Mononucleosomes positioned better before progesterone treatment (left panels) versus after treatment (right panels). (b) Center of well-positioned dinucleosomes.

ChIP-seq datasets. Fragments from each isoform were used as foreground against a combination of the other isoform fragments as background, with a peak defined as having a 3X enrichment over the background, and an FDR < 0.01. As our goal was a of highly-specific peaks for each isoform, the high stringency of the HOMER parameters led to finding relatively few peaks

Isoform	T0 Count	T0 Mean size	T30 Count	T30 Mean size
H1.0	640	151.7	709	151.5
H1.2	1,251	152.2	594	154.4
H1.3	2,433	166.3	2,246	160.4
H1.4	1,159	151.5	536	150.9
H1.5	1,115	151.4	2,044	152.4

Table 4: Competitive peak counts and sizes for each H1 sample.

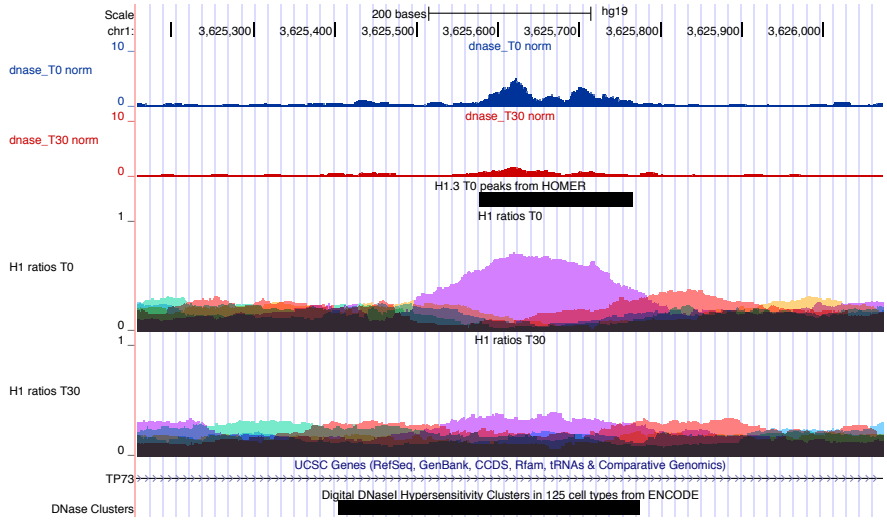


Figure 8: Example peak from the H1.3 dataset. This peak, located within an intron of the TP73 gene on chromosome 1, coincides with DNaseI HS experiments. After 30 min of progesterone the peak is diminished along with the peak of the DNaseI HS track. Shown also is one of the DNaseI HS tracks from ENCODE, where it they have found that this locus is DNaseI hypersensitive in 15 different cell lines, including: another T47D line, as well as MCF-7, LNCaP, Ishikawa cells, which are all happen to be different cancer models for steroid hormone biology.

compared with other ChIP datasets. Peak counts and their mean sizes are listed in Table 4. Visual inspection of various peak regions on the UCSC Genome Browser [247], led to the observation that many H1.3 peaks were coinciding with DNaseI hypersensitivity experiments performed previously [146] (Figure 8). To add to the observations from DNaseI HS sites and H1.3, we gathered various sets of ChIP-seq peaks from other projects in the lab (all with ChIPs performed with and without progesterone), which were

constructed in different ways depending on the ChIP's binding profile (see Table S4). Target proteins included post-translationally modified core histones H3 and H4, the alternate core histone H2A.Z, transcriptional coactivator p300, pioneering factor FOXA1, the insulator protein CTCF, cohesion protein RAD21, RNA Polymerase II (Pol II), and the progesterone receptor (PR). Finally, from Hi-C experiments, we generated a set of 2,031 topologically associating domains [248] (TADs), which are regions of chromatin (usually 1 megabase or larger) where the majority of chromatin interacts with other chromatin within that region. We assigned 1,907 of these TADs one of four classifications (Table 5), which broadly describe the chromatin type within the TAD. Using bedtools [249] to count overlaps between the H1 isoform-specific peaks and the other peaks/regions, and then performing the same overlap test with randomized versions of the H1 peaks (Table S2), we obtained p-values with Fisher exact tests, assessing the association between the H1 peaks and the other features. These p-values are plotted in a heatmap as Figure 9. Several observations from this heatmap stand out: (i) H1.3 peaks associate highly with most of the other genomic features, following the initial observation that they coincided with DNaseI hypersensitivity. (ii) H1.2 peaks, like H1.3 peaks, associate with regions of active promoters/enhancers i.e. H3K4me3/H3K4me1, and both are very unlikely to be found in type III TADs. (iii) many of the peak regions H1.3-specific peaks that are associated with both before and *after* progesterone, are associated with H1.4 or H1.5-specific peaks only gain that association after progesterone. Meanwhile, H1.0-specific peaks remain averse to the ChIP peaks, particularly H2A.Z.

Discussion

At the moment, our dataset offers the highest-resolution study of histone H1 variants, and the first to examine redistribution of the variants in response to a cellular stimulus, in this case progesterone. The overall redundancy of H1 variants from genome-wide correlations suggests that chromatin rarely is in a conformation that favors the binding of one particular variant exclusively, though we did manage to isolate genomic sites that were specific to

TAD	% of TADs	Summarized description
I	14	Small (~ 1 Mbase), gene dense, highly-active genes.
II	38	Small, moderately-active genes.
III	29	Large, low gene density, characterized by bivalent domains: high H3K9me3 and high H3K4me3.
IV	16	Low gene density, low gene expression, high H3K9me3, low in active marks.

Table 5: Topologically-associated domain (TAD) descriptions.

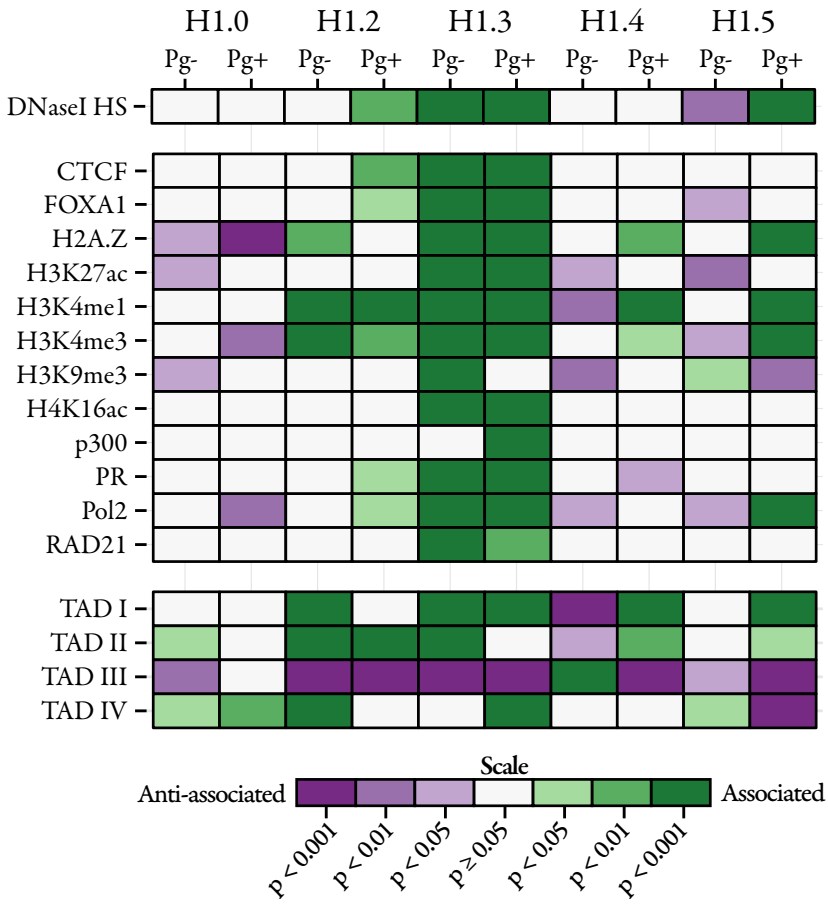


Figure 9: H1 isoform-specific peaks and their associations to other ChIP peaks, DNaseI hypersensitive regions, and topologically-associated domains, based on Fisher Exact Tests of overlap versus expected overlap (see Table S2 for full counts).

one particular variant through isoform-specific peak finding. In the more general regions of chromatin, we observed patterns of deposition that change after progesterone induction. Here we summarize the patterns seen in H1 isoform deposition:

Genes

Surrounding the TSS and TTS sites of all genes lies a region depleted of all isoforms of H1. The extent to which the level of H1 is depleted is directly linked to the expression level of the gene. Genes with higher expression have more depletion of H1. This effect of depletion is less profound in regions surrounding the TTS. Progesterone induction causes a simultaneous increase in H1.5 in TSS/TTS regions and a decrease in H1.3 and H1.2 (mainly downstream of the TSS in H1.2). When comparing sets of genes of similar expression before and after progesterone, the effect of this change in H1.2, H1.3, and H1.5 binding is proportional to the level of expression. This is not the case though when considering sets of genes up or down-regulated by progesterone. Although these genes also follow the pattern of increased H1.5 and decreased H1.3, the level of H1.2 decreases more in up-regulated genes than in down-regulated or non-regulated genes. To varying amounts, the same is true for H1.0 and H1.3, but H1.5 levels increase about the same amount in all three categories, while H1.4 increases, slightly more in down-regulated genes. Although it is somewhat unintuitive to see H1.5, the strongest condenser of chromatin to be so enriched after progesterone, it has however, been previously seen to be associated with active transcription [56,250]. H1.5 has the longest C-terminal domain of the variants. The CTD is rich in lysine residues, contributing to a high positive charge and the increase in residence time on the chromatin [55]. The enrichment may therefore simply be an effect of the difficulty of the removal of H1.5 compared to the other isoforms.

Positioned nucleosomes

The set of well-positioned mononucleosomes are in areas with low H1 binding in general, but the roles of H1.2 and H1.3 would seem to stand out. The high depletion of H1.2 perhaps suggests that has a destabilizing influence on

the positioning of nucleosomes. Looking closer though, while H1.3 does indeed increase slightly in these nucleosomes after they lose their positioning, perhaps more telling is the joint decrease of the H1.0, H1.4, and H1.5 isoforms. The dichotomy between H1.2/H1.3 and H1.0/H1.4/H1.5 has previously been described in the context of chromosome compaction in vitro [50]. Using atomic-force microscopy (ATM), it was shown that H1.0, H1.4, and H1.5 all stabilize chromatin compaction, while H1.2 and H1.3 promote a relaxed chromatin structure. H1.2 was even seen to have a decondensing effect, which is perhaps why it is the most depleted of all.

The positioned dinucleosomes tell a completely different story than the positioned mononucleosomes, and they seem to follow an H1 redistribution pattern similar to the genes. One possibility this is that the well-positioned dinucleosome centers are more gene-associated (exonic) than the well-positioned nucleosome-dyads, although it is true that a dinucleosome occupies a larger region and will overlap with more exons by chance.

Isoform-specific regions

Slight preferences of a certain isoform over another at a single chromatin locus are the norm. That is to say a certain variant may bind there more easily, but other variants can do the job as well. The set of isoform-specific peaks describe regions where the exclusivity to a specific isoform is very high. The question of the mechanism providing exclusivity to one isoform remains. It could be sequence related, a unique conformation of the chromatin, or unique sets of chaperone proteins bound to chromatin nearby, or a combination. We believe that the high number of epigenetic associations suggest proteins are at least partially responsible, but that it is also likely that specific chromatin conformations exist that specifically repel certain isoforms. ChIP experiments can broadly capture these sites, but it is uncertain how long they persist, their underlying mechanism, not to mention the questions of cell-specificity and whether they are the result of gene regulation, or perhaps cause it.

Comparing to similar work

Recently, the DamID method was used to map the binding of H1 variants ge-

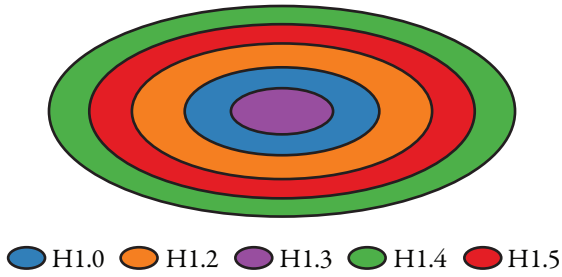


Figure 10: Proposed model of H1 isoform distribution in the T47D nucleus, without progesterone, based mainly on the immunostaining and 100 kb bin analysis, but also the isoform-specific peak locations within TAD types.

nome-wide [52]. As a method, the HA-tag ChIP method offers two particular advantages over DamID for studying H1 variants in a dynamic system: (i) the methyl adenine is only deposited by the fused methyltransferase to GATC sequences roughly every 1 kb, while ChIP offers nucleosome-level resolution, as demonstrated by our 180-200 bp fragment sizes (Figures S3 and S4). Nevertheless, the DamID study found interesting patterns, including depletion of most H1 variants in the promoter regions of genes, increasingly depleted at genes with higher levels of expression. We observed mostly the same pattern, although T47D cells lack the variant H1.1 that the IMR-90 fibroblast lung cells possess, and unfortunately H1.1 was the proverbial "black sheep" in all instances of their comparisons, with the other variants essentially interchangeable. In our case, we find examples of specific enrichment or depletion for each variant, with the most redundant being H1.4. Another recent study has used the same HA:H1 T47D cells [244] that we use for our experiments, with the addition of HA:H1x. When subjecting their data to our analysis pipeline, we found the low genomic coverage of their data to be a bit problematic, but overall their observation that H1.2 and H1.3 are enriched just downstream of TSSs in lowly-expressed genes (without progesterone), while other isoforms are not, is consistent with our results.

Potential models of H1 distribution in the nucleus

The varying localization of H1 isoforms seen by immunostaining and genome-wide visualization of the majority isoform on chromosomes support previous studies involving the localization of chromosomes 4 and 19 in the

nucleus. Chromosome 4, which is gene-poor, is consistently located on the periphery of the nucleus, while by contrast, the gene-rich chromosome 19 is typically in the center of the nucleus [251,252]. Our results from the 100 kb binning analysis indicate only a few¬ of the chromosomes being H1.5-rich: 4, 7, 12, 15, and possibly 3. Many of the H1.3-rich chromosomes are also enriched in H1.0. Likewise, H1.5-rich regions tend to also be enriched in H1.2 or H1.4. Based on these observations, we propose the H1 nucleus model seen in (Figure 10). Despite an overall depletion of H1 at gene transcription start sites, it is possible that central localization of H1.3 could be reason enough that H1.3 is enriched in these regions compared to the other isoforms. A major caveat to the observations concerning H1.3 is that the mass spectrometer showed that it is at low levels in the cell. But retaining these levels from start to finish is a somewhat different analysis. Here, we are continuously examining relative amounts, and we look to find regions where particular isoforms stand out from the others.

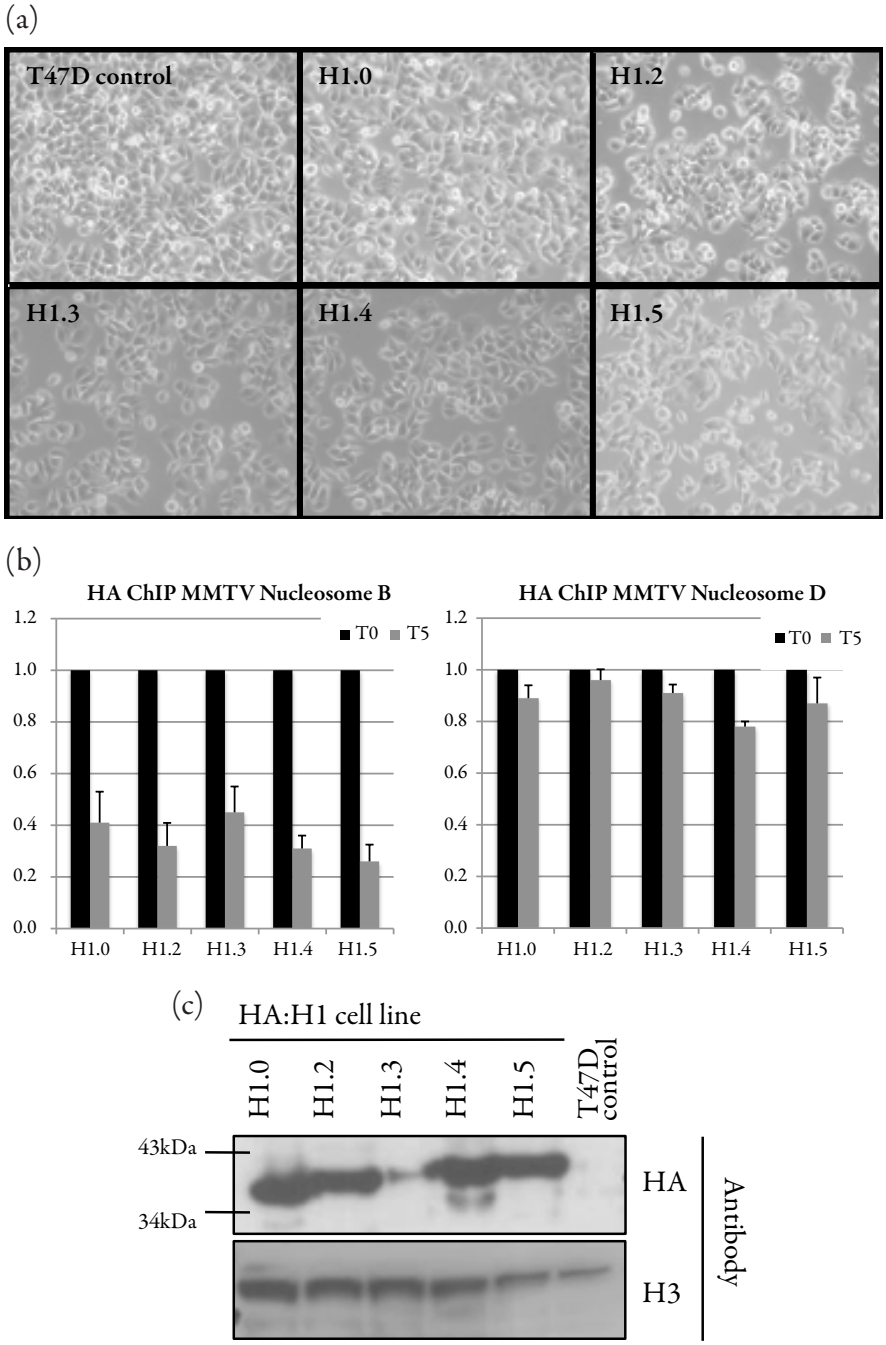


Figure S1: Cell line diagnostics, checking: (a) cell morphology, (b) similar nucleosome displacement (nucleosome B) in MMTV promoter before and after 5 minutes of progesterone treatment, as well as similar non-displacement of nucleosome D. Finally, protein expression was checked by western blot (c).

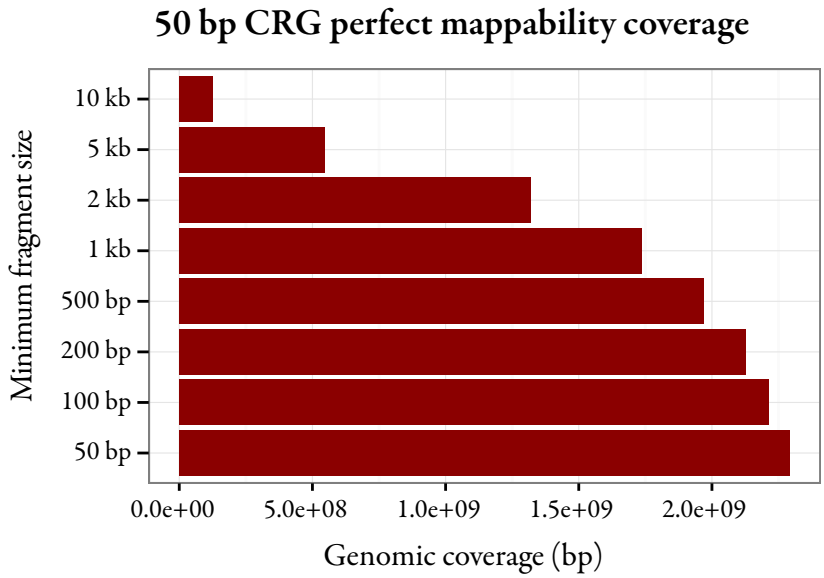


Figure S2: Usable portion of the reference genome (human hg19/GRCh37) when restricting analysis to regions of a minimal size.

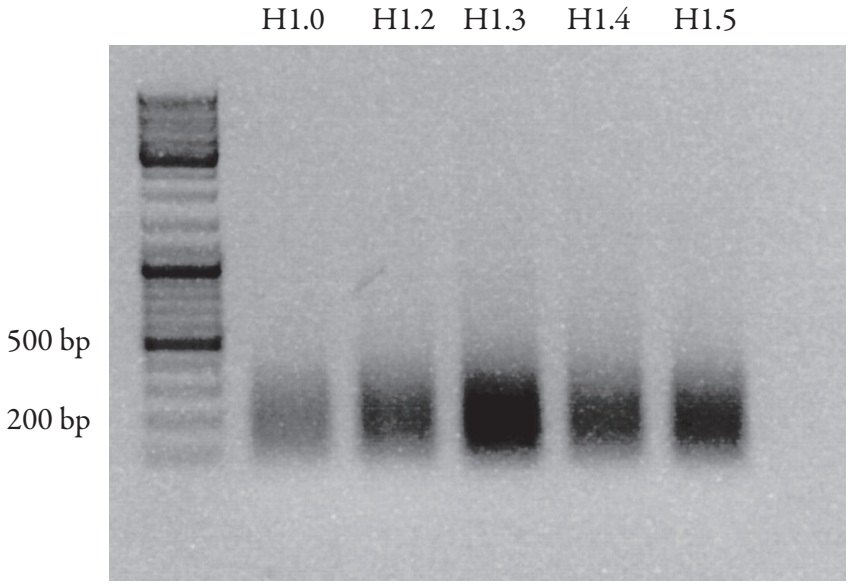


Figure S3: Sizes of sonicated chromatin from untreated HA:H1 ChIP samples. H1.3 has a heavier band because twice as much chromatin was used for this chip, owing to the fact less chromatin is precipicated later in the ChIP from H1.3 samples than the others.

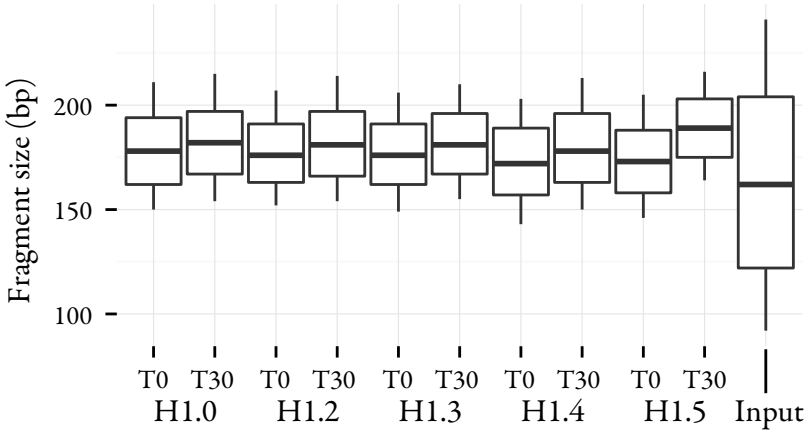


Figure S4: Paired-end sequence fragment size distributions.

	HA:H1.0 T0	HA:H1.2 T0	HA:H1.3 T0	HA:H1.4 T0	HA:H1.5 T0	Input
Sequenced reads	389,582,564	393,420,004	409,290,614	425,477,900	383,288,806	381,611,178
Mapped reads	377,691,736	380,650,445	386,325,981	413,022,141	371,118,656	367,259,197
Paired frags	185,863,667	184,525,455	189,794,505	203,033,468	179,076,478	179,049,136
HDR region	5,159,424	4,989,201	5,615,002	5,948,937	5,038,190	5,081,875
MAPQ < 20	17,227,256	17,669,657	18,186,085	19,151,084	17,853,545	16,145,008
Bad seq. qual.	19,144	20,992	23,046	26,866	17,857	19,156
Duplicates	42,272,426	49,198,072	55,353,046	31,195,000	44,612,225	140,686,454
Reads/frags used	124,072,321	115,351,841	113,781,695	150,132,190	114,364,798	34,427,337

	HA:H1.0 T30	HA:H1.2 T30	HA:H1.3 T30	HA:H1.4 T30	HA:H1.5 T30
Sequenced reads	388,434,080	411,286,430	376,742,002	393,793,876	407,887,012
Mapped reads	377,059,080	399,305,991	354,489,476	381,507,521	392,979,057
Paired frags	185,534,443	196,233,003	173,103,854	185,258,818	180,596,442
HDR region	4,779,704	4,924,517	4,818,189	4,920,018	5,702,900
MAPQ < 20	16,833,079	17,821,988	16,357,429	17,764,591	20,500,081
Bad seq. qual.	19,370	21,722	14,862	18,410	21,627
Duplicates	34,838,623	38,806,067	53,014,719	32,722,754	51,239,945
Reads/frags used	131,644,990	137,271,795	101,559,207	132,504,132	106,366,477

Table S1: Counts from processing the ChIP-sequencing samples.

	# Pg-peaks	mean size	H1.0 T0 (640)		H1.2 T0 (1251)		H1.3 T0 (2433)		H1.4 T0 (1159)		H1.5 T0 (1115)	
			overlap	random	overlap	random	overlap	random	overlap	random	overlap	random
DNaseI HS	66614	533.4	6	14	35	22	585	39	12	20	9	27
CTCF	42390	246.4	1	3	8	12	100	14	10	7	2	5
FOXA1	40384	239.4	1	5	8	9	192	7	8	5	3	11
H2A.Z	19612	2022.4	9	20	36	18	354	31	10	19	10	19
H3K27ac	16837	1411.7	5	14	17	8	299	21	5	15	3	15
H3K4me1	46076	9667.5	102	106	304	186	852	367	132	181	195	180
H3K4me3	43593	1792.4	14	24	63	32	638	62	27	32	24	41
H3K9me3	29566	1,1282.9	62	85	170	142	420	322	117	160	174	136
H4K16ac	23083	706.0	6	6	12	6	220	18	7	7	8	8
P300	442	242.8	0	0	0	0	4	0	0	0	0	0
Pol II	38628	910.3	8	10	29	17	530	39	7	18	11	25
PR	1539	272.6	0	0	0	0	113	2	1	0	0	0
RAD21	22119	186.4	1	0	2	7	53	9	5	2	1	3
TAD I	274	1,085,036.5	48	56	190	107	514	211	63	120	130	115
TAD II	738	1,071,815.7	204	169	444	350	839	698	278	319	339	337
TAD III	571	1,811,558.7	239	282	301	534	592	1009	561	461	377	424
TAD IV	324	1,647,530.9	135	110	292	221	418	429	226	220	243	204

Table S2: H1 isoform-specific peak overlap counts.

	# Pg+ peaks	mean size	H1.0 T30 (640)		H1.2 T30 (1251)		H1.3 T30 (2433)		H1.4 T30 (1159)		H1.5 T30 (1115)	
			overlap	random	overlap	random	overlap	random	overlap	random	overlap	random
DNaseI HS	72521	443.5	4	11	21	7	146	27	10	4	46	19
CTCF	41734	281.2	2	5	11	1	41	14	5	2	8	14
FOXA1	48246	259.4	4	4	12	3	70	16	4	3	13	6
H2A.Z	34625	1823.2	6	26	24	14	132	38	26	9	96	38
H3K27ac	19875	1766.6	8	13	8	9	86	22	9	5	33	27
H3K4me1	55258	10422.3	148	159	164	113	561	450	145	90	651	375
H3K4me3	43047	1731.6	10	25	30	11	174	56	26	13	90	39
H3K9me3	23856	12853.6	72	86	66	76	304	272	70	90	188	242
H4K16ac	29214	649.2	3	8	9	8	45	17	2	5	18	11
p300	6381	317.4	0	3	2	0	25	3	0	1	0	1
Pol II	37921	1056.3	4	16	16	7	111	22	13	6	64	23
PR	48323	315.6	6	8	11	2	99	17	3	10	8	14
RAD21	24481	218.9	0	3	5	1	23	7	5	1	0	4
TAD I	274	1,085,036.5	50	64	77	60	258	183	79	43	410	179
TAD II	738	1,071,815.7	202	201	209	138	645	627	177	138	631	564
TAD III	571	1,811,558.7	263	279	178	268	737	920	157	217	621	830
TAD IV	324	1,647,530.9	175	136	117	103	537	429	108	121	298	390

Table S2: (*continued*).

Isoform	H1.0 T30	H1.2 T30	H1.3 T30	H1.4 T30	H1.5 T30
H1.0 T0	6,402	149	2,685	114	369
H1.2 T0	396	513	92	432	1,482
H1.3 T0	400	127	10,025	144	1,316
H1.4 T0	491	477	368	969	333
H1.5 T0	452	285	253	532	1,547

Table S3: Transitions in highest reads per 100 kb bin among H1 isoforms from before progesterone (rows) to after progesterone (columns). (See figure 3a).

ChIP-seq	Peak-finding method
CTCF	F-seq p<0.01 + Pyicos 1e-5
FOXA1	F-seq p<0.01 + Pyicos 1e-5
H2A.Z	F-seq p<0.01
H3K27ac	F-seq p<0.01
H3K4me1	MACS v2 broadPeak algorithm q<1e-5
H3K4me3	F-seq p<0.01 + Pyicos 1e-5
H3K9me3	BCP
H4K16ac	MACS v2 broadPeak algorithm q<1e-5
p300	F-seq p<0.01 + Pyicos 1e-5
Pol II	F-seq p<0.01 + Pyicos 1e-5
PR	F-seq p<0.01 + Pyicos 1e-5

Table S4: Peak-finding software and parameters used for non-HA:H1 ChIP-seq peaks. Software used included MACS [22], Pyicos [23], F-seq [21], and BCP [24].

Resultats III:

bwtool: a tool for bigWig files

Andy Pohl & Miguel Beato

Bioinformatics (2014) 30 (11): 1618-1619

Pohl A, Beato M. bwtool: a tool for bigWig files.Bioinformatics. 2014 Jun1;30(11):1618-9. doi: 10.1093/bioinformatics/btu056

Abstract

BigWig files are a compressed, indexed, binary format for genome-wide signal data for calculations (e.g. GC percent) or experiments (e.g. ChIP-seq/RNA-seq read depth). bwtool is a tool designed to read bigWig files rapidly and efficiently, providing functionality for extracting data and summarizing it in several ways, globally or at specific regions. Additionally, the tool enables the conversion of the positions of signal data from one genome assembly to another, also known as "lifting". We believe bwtool can be very useful for the analyst frequently working with bigWig data, which is becoming a standard format to represent functional signals along genomes.

Introduction

For many labs it has become an everyday task to generate or to analyze genome-wide data such as ChIP-seq read depth. To facilitate visualization of this data with tools such as the UCSC Genome Browser [254] or ENSEMBL [255], or for further processing, it is common to use the wiggle (WIG) file format. This format is not without a few disadvantages, principally that the files can become quite large, particularly when care is not taken to store the data at a minimally-necessary decimal precision. Another disadvantage is that wiggles exist in three different forms, the choice of which depends on the sparseness of the data. Programs that expect WIG data do not always allow all three formats interchangeably.

The bigWig format [256] was created as a means for the UCSC Genome Browser to access real-valued signal data remotely hosted on HTTP/FTP servers worldwide. The format is binary, compressed, indexed, and allows random access to directly query a subset of the larger dataset. In general, programs designed to read bigWig files should treat remote URLs of bigWigs the same as if they were local to that computer. bigWig uses an indexing strategy similar to other binary/indexed formats such as bigBed [256], BAM [178], and tabix-based formats [257], but unlike BAM or tabix-based formats, bigWig is specific to numerical data. WIG and BAM are both common data formats and are utilized by many applications, e.g. MACS [195] and MISO [258] respectively, but to date there are not many applications that accommodate bigWig data.

We have created command-line software under the UNIX operating system called bwtool in a similar spirit to bedtools [249] or samtools [178] that offers the possibility to carry out a number of diverse operations on bigWigs in a convenient way. Until now, the common procedure to access the data within bigWig files has been to use the tools available from UCSC: bigWigToWig, bigWigSummary, bigWigAverageOverBed, bigWigMerge, bigWigCorrelate, or bigWigInfo. These offer some basic usability for bigWigs. bigWigInfo provides instant information about a bigWig file and is useful for glancing at the overall mean and standard deviation as well as seeing how

many bases are covered by the signal. bigWigToWig is indispensible as it is occasionally necessary to convert a bigWig into the original WIG to utilize legacy software. Beyond those two, bwtool provides additional features and flexibility not found in other software.

Description

The bwtool program is designed to rapidly collect summary statistics and do common wiggle manipulations. The program is actually a collection of utilities (the names of which are in bold), which allow for the following features:

- **Aggregate** data by averaging it over a series of given intervals with respect to central bases. This common aggregation procedure is used to produce plots showing enrichment, but has a tendency to be problematic, particularly when centering on genomic features without a known strand or directionality [209]. For this reason, simple k-means functionality is built-in to group regions with similar profiles. Figure 1 demonstrates the aggregate program on data collected from the ENCODE project [219].

- "**Lift**", i.e. project data from one genome assembly to another using a "liftOver chain" file, available from the UCSC Genome Browser Utilities Page [259]. Lifting data often results in a small percentage of lost data, so care must be taken to ensure that the only lifted data analyzed is that which is within regions lifting correctly. Options are available to catalog all of the problematic regions involved.

- Quickly **find** regions in the bigWig exhibiting local minima/maxima, or above/below specified thresholds.

- Extract equally sized intervals of data as a **matrix** or as a sliding **window** at adjustable steps and sizes. Again, clustering is available as an option when extracting data as a matrix. A **random** matrix of data can also be produced, with the ability to exclude specific regions in the genome. Unequally sized intervals can be also extracted with the **extract** utility.

- Another way to extract data from multiple bigWigs is to use the **paste**

utility. This outputs tab-delimited data from a set of bigWigs, one base per line. Pasting bigWigs together makes it possible to perform many complex calculations with small auxiliary scripts. In this way, the functionality of bwtool can easily expand upon the functionality of bigWigMerge and bigWigCorrelate from UCSC.

- Discretize the real-valued signal into letters, using the **SAX** algorithm [260].
- **Removing** data based on thresholds and specific regions if desired. Conversely, regions missing data in a bigWig can be replaced with a constant using the **fill** utility.
- **Summarize** data at specific regions. This functionality is similar to the combined programs of bigWigSummary and bigWigAverageOverBed, with the addition of median and optional quantile information in the output.

Common options to many of the features include the ability to specify the decimal precision, to fill missing bases with a given value, or to provide a bed file specifying specific regions of the bigWig to read.

Usage and Availability

bwtool is command-line software for UNIX, a common platform for bioinformatics researchers to conduct analysis. Running the bwtool command without additional parameters displays a description of the various utilities and some general options. Combined with a utility name, bwtool will display specific information about how to perform an operation using that utility. A detailed guide has been created on bwtool's web page (http://cromatina.crg.eu/bwtool) to provide thorough examples of using the program.

bwtool is written in C. The source code for the program is available on its GitHub web page. Distributed (with permission) with bwtool is the basic C library from Jim Kent that is needed for routines specific to bigWig data, as well as other algorithmic code. He and the University of California hold the copyright to this specific library, but the remaining code is covered

by the GNU Public License v3. bwtools makes use of GNU autotools to simplify the installation process to the standard "./configure", "make", "make install" procedure most UNIX users will be familiar with. To verify the accuracy of the software, tests may be run with "make check". bwtool does not require additional libraries that are not typically found in common UNIX environments, but if the GNU Scientific Library is installed, it will make use of that for the **random** utility.

Acknowledgements

Thanks to Daniel Soronellas, João Curado, Alessandra Breschi, Roderic Guigó, Jakob Skou Pedersen, Jim Kent, and Brian Raney for testing the program and providing feedback and advice prior to release.

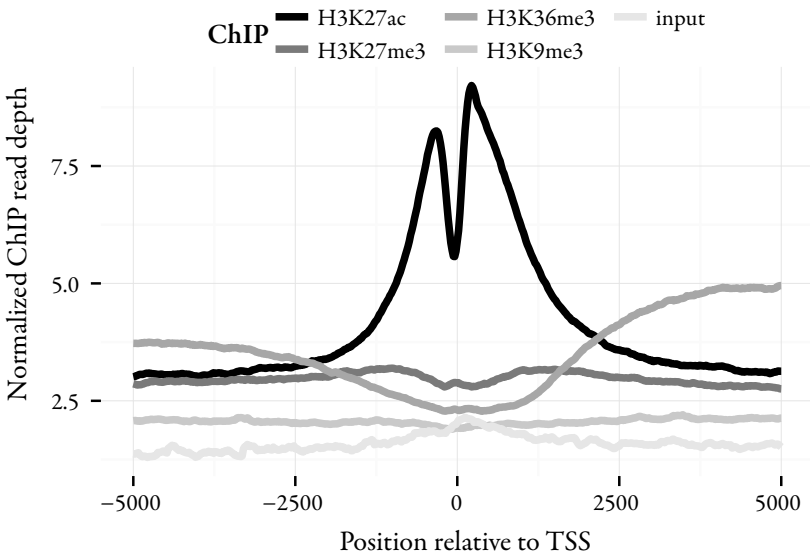


Figure 1: Example of aggregated plots of different histone modification ChIP sequence read-depth signals from MCF7 cells from ENCODE aligned at each of the 20,330 protein-coding gene tran-scription start sites in GENCODE release v17 (Harrow *et al.*, 2012) [278]. See supplement for instructions on how to reproduce this plot. The raw signals in this example are not normalized, so spe-cific values cannot be compared between signals, however the morphological differences in averaged profiles are nevertheless useful in characterizing the patterns of each histone mark.

Discussió

Reoccurring themes

Both the nucleosome and the histone H1 projects I have presented share many similarities. They begin with breast cancer cells on a plate in an incubator, serum-starved and arrested in G0/G1 phase prior to a hormone stimulus. They end with billions of sequenced nucleotides from the eventual isolation of the DNA of those very cells. The steps in between vary depending on the experiment, but several challenges remain constant regardless.

The karyotype of polyploidy/aneuploid genomes like T47D's must be handled properly from the beginning, otherwise differential gene expression or differential ChIP binding may be masked by allelic differences. The usual strategy to working with high-throughput sequence-based experiments in human is to select more diploid cell types such as SV40-immortalized fibroblasts or keratinocytes [261], lymphoblastoids, or sometimes primary cells, particularly leukocytes. Pertinent examples include the original nucleosome positioning studies in humans [115,262], or the histone H1 DamID study [52]. Through trial and error we settled on a karyotype normalization algorithm that is robust not just to varying fragment coverage among datasets, but also varying fragment sizes like those from paired-end sequencing. Varying fragment sizes is not normally accounted for in fragment depth normalization of genomic samples, ENCODE's Wiggler software uses merely the fragment counts [263]. We have leveraged this algorithm to make use of aneuploid cell models we need for the investigation of steroid hormone-based gene regulation and chromatin dynamics.

Apart from karyotype another problem common to these projects is the ubiquitous nature of the data. Unlike transcription factors, for example, histones and nucleosomes are bound all across the genome. In these cases, local depletion is often easier to explain than enrichment. These types of data

are expensive to produce, because they require a sequence coverage level approaching that of genome assembly. One way to circumvent the need for such a high level of coverage is to use paired-end sequencing. Not only is more of the genome mappable with paired-end data, but also knowing the end-point of a fragment substantially improves the ability to position nucleosomes at base pair resolution. Consider two MNase-seq reads with staggered genomic alignments with 5' ends separated by 5 bp. While a single-end experiment would identify two separate nucleosome positions, it is possible, and often seen with paired-end sequencing, that one fragment is simply larger than the other, and superimposed on the genome they share the same midpoint base and therefore only account for a single nucleosome position. Thus the only way to achieve a high-degree of nucleosome positioning with single-end data is to have both an extremely uniform size of fragments, and to know that size so the fragments are extended properly. Having one of those criteria and not the other will ultimately lead to inferior positioning. Paired-end sequencing requires neither for good positioning, and because our data make use of higher fragment sizes associated with the inclusion or partial inclusion of histone H1, we also expect a broader range of fragment sizes.

With such a high degree of background in high-density ChIP-seq data such as that from our H1 project, peak-finders, which are usually designed for more isolated peaks, become unreliable. We were able to isolate variant-specific peaks for each isoform in our H1 study using a variant ChIP-seq as foreground, with the combination of the other variant ChIP-seqs as background. This is not what HOMER was designed for, but nevertheless the peaks we obtained seem genuine, and although we have not associated any with a regulatory function, they coincide with known epigenetic markers, in a variant-specific manner, with high degrees of statistical significance.

Although certain challenges present themselves when studying the kind of data we collect in the kind of cells we culture, we are confident these issues can always be circumvented with careful approaches or using the more sensitive assay.

Final Considerations

Our conclusion that partial or unstable nucleosomes occupy the region known as the "nucleosome-free region" is based on our own observations, but also studies [230,234], that are either unnoticed or consciously ignored.

If we are at a point where nucleosomal DNA recovered in the nucleosome-free region using protection assays like MNase-seq must be called "non-nucleosome bound" DNA, simply because of where in the genome they are found, and their size [238], then the concept itself needs to be revisited. MNase has typically been used in high concentrations and that is one cause of the depletion in these regions, but just as salt concentration is a factor for higher-order chromatin compaction [32], it has also been seen to affect the stability of H2A.Z/H3.3 nucleosomes [232], which are likely the dominant form in the NFR.

I think it is perfectly natural to expect resistance to our interpretation of our observed MNase-seq data. At the moment, it still needs to be reinforced by demonstrating the effect of several different concentrations of MNase on the zero nucleosome. Additionally, we want to know more about how the knockdown of Brg1/Brm proteins stabilizes nucleosome zero. Ultimately, our goal is to find an inherent property of the promoter sequence or structure that gives rise to the nucleosome zero in the first place. It is our hypothesis that DNA sequence exists that specifically evicts H2A/H2B dimers from nucleosomes when slid into place by certain chromatin remodelers. Our observations of the zero nucleosome point to the presence of nucleosomes with proportionally lower H2A content, and shorter sequenced fragments due to a smaller-than-canonical nucleosome protecting the DNA from cleavage by MNase.

Tying it all together

Logically speaking, the deposition of H1 should have an influence on nucleosome positioning, and vice versa. H1 fundamentally alters the structure of chromatin and requires an additional 20 bp of linker DNA. A recent article

described the 30 nm chromatin fiber structure using Cryo-EM, of in vitro reconstituted chromatin with histone H1.4 and nucleosome repeat lengths of 177 or 187 bp [66] (see Figure 1). They found chromatin organized in tetranucleosomal units, with H1 asymmetrically bound to the nucleosome core particle, contacting the entering and exiting DNA as well as the nucleosome dyad. Within the tetranucleosome, dinucleosomal stacks are formed from alternating as opposed to adjacent mononucleosomes in a zig-zag conformation. These dinucleosome stacks are separated by the linker DNA and twisted in a ¼ turn such that repeated tetranucleosomes form a left-handed double-helical structure. A metaphor for the asymmetry of H1 in this model describes the nucleosome as a coin, the nucleosome wrapped around the edges by DNA, with the "head" of the coined having a smaller portion of the H1 bound than the "tail". Dinucleosome stacks are then configured in a tail-tail manner, with H1 on more on the exterior of the tetranucleosome. The authors acknowledge that these observations are made under very controlled conditions and the environment in vivo is probably more variable and dynamic, but the concept is interesting and it proposes new questions for our H1 study. Part of the redundancy we see might actually be a case of H1 crosslinked across the dinucleosome stack at a nucleosome that is actually a nucleosome further away from the one we believe it is. And whether or not this is the case, is it possible to detect tetranucleosomes, or at least alternating dinucleosome stacks based on the asymmetrical configuration of H1? We have seen positioned dinucleosomes in regulatory regions surrounding gene transcription start sites with a tendency to phase at a peridiocity similar to mononucleosomes. This type of phasing is probably more compatible with a zigzag 30 nm chromatin fiber. In regions where the dinculeosome phasing periodicity seems to form dinucleosome units, the solenoid model, with interacting adjacent mononucleosomes, is a more likely fit.

Nucleosomes have been positioned in yeast with much better precision than in humans. With a genome 300 times larger, and cells that are usually more difficult to culture in synchronized, homogenous conditions, positioning nucleosomes in humans is much more difficult simply for technical reasons. Biologically speaking though, the main difference in the chromatin

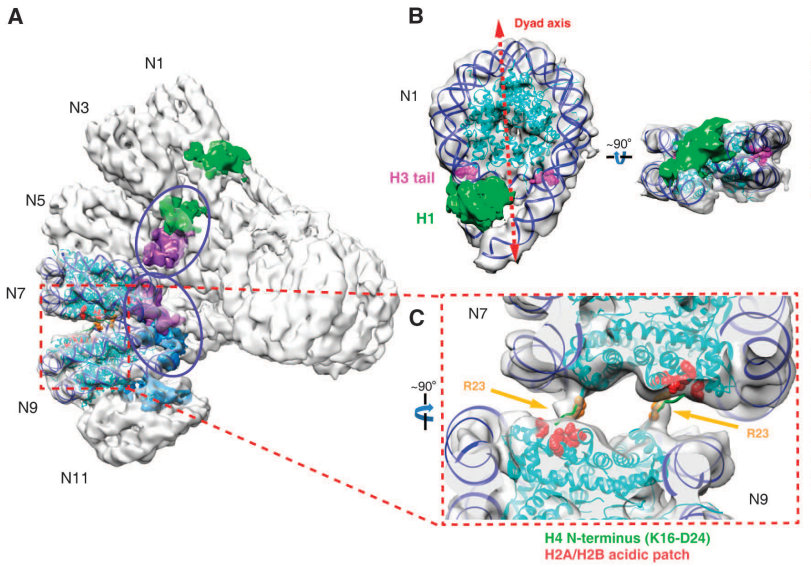


Figure 1: (From Song, et. al (2014) [66]). Cryo-EM structure of 30 nm chromatin fiber showing stacked tetranucleosomes (A) showing position of H1 (green) in relation to the other nucleosomes. (B) shows the asymmetry of H1's binding in more detail.

of the two species is the lack of histone H1 in yeast chromatin and the associated much shorter linker DNA. But the interpretation should not be that H1 contributes to poor positioning. Without H1, and with shorter linker lengths, there are simply fewer positions for the yeast nucleosomes to occupy. H1 is normally understood to stabilize nucleosomes. Therefore we find it a bit surprising that the nucleosomes we found with the best positioning scores were both depleted of H1 (particularly H1.2) and were flanked with long linkers. For this reason, we think that our nucleosome scoring algorithm should include consideration of the flanking linker DNA sequences.

We previously found that the symmetry of curvature in DNA had reasonable predictive power for finding nucleosome positions in yeast [264]. Although one of my stated objectives was to find additional links between structure and nucleosome positioning, more specifically, the goal was to improve upon this method and apply it to the human genome, using the longer linkers and H1 binding as an advantage. Unfortunately I never managed to improve the SymCurv algorithm so that it would perform well in the human genome. I would suspect that the asymmetrical binding of H1 directly

thwarts this.

Histone H1 and nucleosome positioning are linked topics, but are usually not treated as such, including in our case. An easy way would be to combine the two methods by treating the chromatin with MNase instead of shearing the chromatin with sonication prior to immunoprecipitation. Even better would be to trim the DNA down to what is only H1-bound in a method similar to ChIP-exo [265], which has been used to find transcription factor binding sites at single base-pair resolution. In either case though, we are still limited by antibody methods. Although we can discriminate between H1 variant binding using HA-tagged variants, we still do so using separate plates in what are effectively separate experiments. It would be nicer to have the full spectrum of H1 and core histone modifications known, at all sites in the genome, from the same population of cells, perhaps synchronized and sorted by FACS using cell cycle markers. Or have that have the full spectrum of histones assayed using a single human cell. But for this, the methods and technology are not available yet.

The anti-antibody future of chromatin biology and epigenetics

In the last decade, many advances have been made across whole sectors of biotechnology increasing the resolution of microscopes, the resolution of mass spectrometers, or the throughput of sequencing. Although each of these technologies greatly benefits the study of epigenetics, the sector that stands to help epigenetics the most could be the continually evolving technology associated with proteomics. Antibody-based assays like ChIP have become extremely useful, but have varying success rates, depending on the antibody. Raising antibodies generated from animals exposed to specific antigens is time-consuming and expensive. Ironically, though antibodies are prized for their specificity, in a way it is one of their most limiting factors. To properly capture the dynamics of the state of chromatin with ChIP, a variety of antibodies are needed for what are usually separate experiments, which then need multiple replicates, multiplied again by however many chromatin states are intended to be studied in the dynamic system.

The "holy grail" solution to this limitation is the mass-spectrometry of chromatin-binding proteins isolated from specific loci. Pioneering attempts have been made with this type of experiment with PICh [266], or ChAP-MS [267], but at this time are still far from perfect. PICh requires a lot of material, so is generally restricted to highly repetitive regions such as telomeres. ChAP-MS can be made specific to any region, but the requirement of the endogenous insertion of a LexA binding site upstream of the locus of interest precludes the method from high-throughput analysis.

A common method to reduce bias in antibodies is to ectopically express tagged versions of the proteins of interest, and use an antibody against those tags instead. Common tags are HA, FLAG, or MYC. Florescent proteins like GFP are also used, but are bulkier and are more likely to perturb endogenous protein function [268]. Perhaps the biggest problem with using tagged proteins is that even when care is taken to reduce the overexpression of the protein by cloning the promoter region in addition to the gene's sequence, e.g. BAC-based recombineering [269], the protein of interest remains overexpressed. Proteins may be endogenously tagged using TALENs (transcription activator-like effector nucleases) [270], or their predecessors, Zinc finger nucleases [271,272], but these techniques require the generation of custom proteins, which is difficult and tedious. These methods are also often not perfectly specific to the sequence they are supposed to cut, resulting in what are known as "off target" double-strand breaks. Too many of these may end up overwhelming the DNA repair system of the cell, eventually causing toxicity and even cell death.

Recently, the CRISPR/Cas system has been introduced [273-276], combining the ease of designing siRNAs and the power of TALENs. CRISPR is also capable of editing genomes, and fully knocking-out genes, but all that is necessary is to design RNA guides. If desired, various proteins can also be fused to Cas and directed to specific genomic sites by CRISPR.

Putting these two pieces together: site-specific proteomics and simpler genomic engineering, could be the key to a bright future without such a heavy reliance on antibodies. A screen of Cas/biotin CRISPR constructs directed to as many loci as possible genome-wide, with each sample's chroma-

tin sonicated and pulled down with streptavidin, and analyzed by mass spec, would simplify the process of studying chromatin. At the moment, such a technique is probably mainly limited by the mass spectrometer, although they continue to increase in sensitivity. A genome-wide CRISPICh technique would be a drop-in replacement for the H1 study presented here, and would have the advantage that all of the other chromatin-bound proteins and their modifications could be analyzed as well.

Making computational accommodations

The data accompanying technological advances, particularly sequencing data, is growing at an alarming rate. Increasing raw data leads to increased need for additional and expanded computational resources. It may be the case when sequencers are as common as other bench-top equipment such as centrifuges or shakers, biologists will also discard sequence data more often, just as they do with bad gels or bad exposures. But what is more likely the case is that the biologist will generate sequence data and struggle to find somewhere to store it. While it was once the case that an analyst could get by with a powerful workstation, bioinformaticians are becoming increasingly dependent on powerful computer clusters, and even making use of cloud computing. The brutal irony of it all is that more powerful computers offer developers excuses to write worse software, mitigating to a large extent technological advances made on the hardware end. Fortunately, there are good examples of powerful software in the field of bioinformatics, such as samtools, that allow us to stay ahead of the curve. bwtool was written with power in mind, but also to have an intuitive interface, a simple installation procedure, and good online documentation in the form of a wiki (http://cromatina.crg.eu/bwtool). Its use of compressed bigWig files also aids in the unfortunate but natural excessive use of shared data storage.

Conclusions

1. bwtool has been developed to utilize the common bigWig format, directly, in a number of convenient ways.
2. A nucleosome occupying the NFR (centered at 25 bp upstream of the TSS) has been observed in MNase nucleosome positioning experiments with a gentle digestion of MNase.
3. The NFR is depleted of nucleosomes in chromatin, but also free DNA of the NFR is hypersensitive to MNase digestion.
4. The nucleosome positioning score reveals a "zero nucleosome" even in data sets with a strong nucleosome depletion in the NFR.
5. The occupancy of the "zero nucleosome" correlates with binding of histone H2A.Z, while at the same time the NFR is depleted of histone H2A.
6. Paired-end reads from the NFR region are shorter than the reads in the flanking -1 and +1 nucleosomes.
7. All of the histone H1 variants are depleted in the TSS, with H1.3 being the lest depleted variant followed by H1.2.
8. Although proportional differences in binding amongst H1 variants are slight at a given 100 kb-sized locus, loci with a higher proportion of a certain variant cluster into large chromosomal regions.
9. After progesterone induction there is widespread enrichment of H1.5 throughout gene bodies.
10. H1.2 is more depleted than the others at well-positioned nucleosomes flanked by longer linkers.

11. H1.3 highly enriched nucleosomes correlate with epigenetic marks associated with gene activation.
12. Levels of H1.2 binding are highly-enriched across the bodies of the genes we previously observed as down-regulated after depleting H1.2.

Bibliografia

[1] Jacob F, Monod J. Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol (1961), 3:318–356.

[2] Duclaux E. Traité de microbiologie. (1899).

[3] Clever U, Karlson P. Induction of puff changes in the salivary gland chromosomes of Chironomus tentans by ecdysone. Exp Cell Res (1960), 20:623–626.

[4] Lifton RP, Goldberg ML, Karp RW, Hogness DS. The organization of the histone genes in Drosophila melanogaster: functional and evolutionary implications. Cold Spring Harb Symp Quant Biol (1978), 42 Pt 2:1047–1051.

[5] Banerji J, Rusconi S, Schaffner W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. Cell (1981), 27:299–308.

[6] Kornberg RD. Chromatin structure: a repeating unit of histones and DNA. Science (1974), 184:868–871.

[7] Allfrey VG, Faulkner R, Mirsky AE. Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. Proc Natl Acad Sci USA (1964), 51:786–794.

[8] Allfrey VG, Mirsky AE. Structural Modifications of Histones and their Possible Role in the Regulation of RNA Synthesis. Science (1964), 144:559–559.

[9] Brownell JE, Zhou J, Ranalli T, Kobayashi R, Edmondson DG, Roth SY, Allis CD. Tetrahymena histone acetyltransferase A: a homolog to yeast Gcn5p linking histone acetylation to gene activation. Cell (1996), 84:843–851.

[10] Taunton J, Hassig CA, Schreiber SL. A mammalian histone deacetylase related to the yeast transcriptional regulator Rpd3p. Science (1996), 272:408–411.

[11] Chow LT, Gelinas RE, Broker TR, Roberts RJ. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. Cell

(1977), 12:1–8.

[12] Berget SM, Moore C, Sharp PA. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. Proc Natl Acad Sci USA (1977), 74:3171–3175.

[13] Narlikar GJ, Fan H-Y, Kingston RE. Cooperation between complexes that regulate chromatin structure and transcription. Cell (2002), 108:475–487.

[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), 389:251–260.

[15] Tremethick DJ. Higher-Order Structures of Chromatin: The Elusive 30 nm Fiber. Cell (2007), 128:651–654.

[16] Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature (2012), 485:376–380.

[17] Paweletz N. Walther Flemming: pioneer of mitosis research. Nat Rev Mol Cell Biol (2001), 2:72–75.

[18] Chen X, Chen Z, Chen H, Su Z, Yang J, Lin F, Shi S, He X. Nucleosomes suppress spontaneous mutations base-specifically in eukaryotes. Science (2012), 335:1235–1238.

[19] White CL, Suto RK, Luger K. Structure of the yeast nucleosome core particle reveals fundamental changes in internucleosome interactions. EMBO J (2001), 20:5207–5218.

[20] Lorch Y, Zhang M, Kornberg RD. Histone octamer transfer by a chromatin-remodeling complex. Cell (1999), 96:389–392.

[21] Bradbury EM. Nucleosome and chromatin structures and functions. J Cell Biochem Suppl (1998), 30-31:177–184.

[22] Margueron R, Trojer P, Reinberg D. The key to development: interpreting the histone code? Current Opinion in Genetics & Development (2005), 15:163–176.

[23] Jenuwein T, Allis CD. Translating the histone code. Science (2001), 293:1074–1080.

[24] Akhmanova A, Miedema K, Hennig W. Identification and character-

ization of the Drosophila histone H4 replacement gene. FEBS Lett (1996), 388:219–222.

[25] Hernández-Muñoz I, Lund AH, van der Stoop P, Boutsma E, Muijrers I, Verhoeven E, Nusinow DA, Panning B, Marahrens Y, van Lohuizen M. Stable X chromosome inactivation involves the PRC1 Polycomb complex and requires histone MACROH2A1 and the CULLIN3/SPOP ubiquitin E3 ligase. Proc Natl Acad Sci USA (2005), 102:7635–7640.

[26] Buschbeck M, Uribesalgo I, Wibowo I, Rué P, Martin D, Gutierrez A, Morey L, Guigó R, López-Schier H, Di Croce L. The histone variant macroH2A is an epigenetic regulator of key developmental genes. Nat Struct Mol Biol (2009), 16:1074–1079.

[27] Kraushaar DC, Jin W, Maunakea A, Abraham B, Ha M, Zhao K. Genome-wide incorporation dynamics reveal distinct categories of turnover for the histone variant H3.3. Genome Biol (2013), 14:R121.

[28] Xu Y, Ayrapetov MK, Xu C, Gursoy-Yuzugullu O, Hu Y, Price BD. Histone H2A.Z controls a critical chromatin remodeling step required for DNA double-strand break repair. Mol Cell (2012), 48:723–733.

[29] Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. High-resolution profiling of histone methylations in the human genome. Cell (2007), 129:823–837.

[30] Bustin M, Catez F, Lim J-H. The dynamics of histone H1 function in chromatin. Mol Cell (2005), 17:617–620.

[31] Bryant JM, Govin J, Zhang L, Donahue G, Pugh BF, Berger SL. The linker histone plays a dual role during gametogenesis in Saccharomyces cerevisiae. Mol Cell Biol (2012), 32:2771–2783.

[32] Thoma F, Koller T, Klug A. Involvement of histone H1 in the organization of the nucleosome and of the salt-dependent superstructures of chromatin. The Journal of Cell Biology (1979), 83:403–427.

[33] Hendzel MJ, Lever MA, Crawford E, Th'ng JPH. The C-terminal domain is the primary determinant of histone H1 binding to chromatin in vivo. J Biol Chem (2004), 279:20028–20034.

[34] Hill DA. Influence of linker histone H1 on chromatin remodeling.

Biochemistry and Cell Biology (2001).

[35] Fyodorov DV, Kadonaga JT. Chromatin assembly in vitro with purified recombinant ACF and NAP-1. Meth Enzymol (2003), 371:499–515.

[36] Olins DE. Interaction of Lysine-rich Histones and DNA. J Mol Biol (1969), 43:439–460.

[37] Roth SY, Allis CD. Chromatin condensation: does histone H1 dephosphorylation play a role? Trends Biochem Sci (1992), 17:93–98.

[38] Weiss T, Hergeth S, Zeissler U, Izzo A, Tropberger P, Zee BM, Dundr M, Garcia BA, Daujat S, Schneider R. Histone H1 variant-specific lysine methylation by G9a/KMT1C and Glp1/KMT1D. Epigenetics Chromatin (2010), 3:7.

[39] Christophorou MA, Castelo-Branco G, Halley-Stott RP, Oliveira CS, Loos R, Radzisheuskaya A, Mowen KA, Bertone P, Silva JCR, Zernicka-Goetz M, Nielsen ML, Gurdon JB, Kouzarides T. Citrullination regulates pluripotency and histone H1 binding to chromatin. Nature (2014), 507:104–108.

[40] Kinkade JM, Cole RD. A structural comparison of different lysine-rich histones of calf thymus. J Biol Chem (1966), 241:5798–5805.

[41] Wierzbicki AT, Jerzmanowski A. Suppression of histone H1 genes in Arabidopsis results in heritable developmental defects and stochastic changes in DNA methylation. Genetics (2005), 169:997–1008.

[42] Przewloka MR, Wierzbicki AT, Slusarczyk J, Kuraś M, Grasser KD, Stemmer C, Jerzmanowski A. The "drought-inducible" histone H1s of tobacco play no role in male sterility linked to alterations in H1 variants. Planta (2002), 215:371–379.

[43] Takami Y, Nishi R, Nakayama T. Histone H1 variants play individual roles in transcription regulation in the DT40 chicken B cell line. Biochemical and Biophysical Research Communications (2000), 268:501–508.

[44] Fan Y, Nikitina T, Morin-Kensicki EM, Zhao J, Magnuson TR, Woodcock CL, Skoultchi AI. H1 linker histones are essential for mouse development and affect nucleosome spacing in vivo. Mol Cell Biol

(2003), 23:4559–4572.

[45] Fan Y, Sirotkin A, Russell RG, Ayala J, Skoultchi AI. Individual somatic H1 subtypes are dispensable for mouse development even in mice lacking the H1(0) replacement subtype. Mol Cell Biol (2001), 21:7933–7943.

[46] Schulze E, Schulze B. The vertebrate linker histones H1 zero, H5, and H1M are descendants of invertebrate "orphon" histone H1 genes. J Mol Evol (1995), 41:833–840.

[47] Roche J, Gorka C, Goeltz P, Lawrence JJ. Association of histone H1(0) with a gene repressed during liver development. Nature (1985), 314:197–198.

[48] Doenecke D, Tönjes R. Differential distribution of lysine and arginine residues in the closely related histones H1 and H5. Analysis of a human H1 gene. J Mol Biol (1986), 187:461–464.

[49] Burfeind P, Hoyer-Fender S, Doenecke D, Hochhuth C, Engel W. Expression and chromosomal mapping of the gene encoding the human histone H1.1. Hum Genet (1994), 94:633–639.

[50] Clausell J, Happel N, Hale TK, Doenecke D, Beato M. Histone H1 subtypes differentially modulate chromatin condensation without preventing ATP-dependent remodeling by SWI/SNF or NURF. PLoS ONE (2009), 4:e0007243.

[51] Talasz H, Sapojnikova N, Helliger W, Lindner H, Puschendorf B. In vitro binding of H1 histone subtypes to nucleosomal organized mouse mammary tumor virus long terminal repeat promotor. J Biol Chem (1998), 273:32236–32243.

[52] Izzo A, Kamieniarz-Gdula K, Ramírez F, Noureen N, Kind J, Manke T, van Steensel B, Schneider R. The Genomic Landscape of the Somatic Linker Histone Subtypes H1.1 to H1.5 in Human Cells. CellReports (2013), 3:2142–2154.

[53] Meergans T, Albig W, Doenecke D. Varied expression patterns of human H1 histone genes in different cell lines. DNA Cell Biol (1997), 16:1041–1049.

[54] Konishi A, Shimizu S, Hirota J, Takao T, Fan Y, Matsuoka Y, Zhang

L, Yoneda Y, Fujii Y, Skoultchi AI, Tsujimoto Y. Involvement of histone H1.2 in apoptosis induced by DNA double-strand breaks. Cell (2003), 114:673–688.

[55] Th'ng JPH, Sung R, Ye M, Hendzel MJ. H1 family histones in the nucleus. Control of binding and localization by the C-terminal domain. J Biol Chem (2005), 280:27809–27814.

[56] Parseghian MH, Newcomb RL, Winokur ST, Hamkalo BA. The distribution of somatic H1 subtypes is non-random on active vs. inactive chromatin: distribution in human fetal fibroblasts. Chromosome Res (2000), 8:405–424.

[57] Sancho M, Diani E, Beato M, Jordan A. Depletion of human histone H1 variants uncovers specific roles in gene expression and cell growth. PLoS Genet (2008), 4:e1000227.

[58] Li J-Y, Patterson M, Mikkola HKA, Lowry WE, Kurdistani SK. Dynamic distribution of linker histone H1.5 in cellular differentiation. PLoS Genet (2012), 8:e1002879.

[59] Terme J-M, Sesé B, Millán-Ariño L, Mayor R, Izpisúa Belmonte JC, Barrero MJ, Jordan A. Histone H1 variants are differentially expressed and incorporated into chromatin during differentiation and reprogramming to pluripotency. Journal of Biological Chemistry (2011), 286:35347–35357.

[60] Rubin AF, Green P. Comment on "The consensus coding sequences of human breast and colorectal cancers". Science (2007), 317:1500–1500.

[61] Stoldt S, Wenzel D, Schulze E, Doenecke D, Happel N. G1 phase-dependent nucleolar accumulation of human histone H1x. Biol Cell (2007), 99:541–552.

[62] Warneboldt J, Haller F, Horstmann O, Danner BC, Füzesi L, Doenecke D, Happel N. Histone H1x is highly expressed in human neuroendocrine cells and tumours. BMC Cancer (2008), 8:388.

[63] Quénet D, McNally JG, Dalal Y. Through thick and thin: the conundrum of chromatin fibre folding in vivo. EMBO Rep (2012), 13:943–944.

[64] Fussner E, Strauss M, Djuric U, Li R, Ahmed K, Hart M, Ellis J, Bazett-Jones DP. Open and closed domains in the mouse genome are configured as 10-nm chromatin fibres. EMBO Rep (2012), 13:992–996.

[65] Joti Y, Hikima T, Nishino Y, Kamada F, Hihara S, Takata H, Ishikawa T, Maeshima K. Chromosomes without a 30-nm chromatin fiber. Nucleus (2012), 3:404–410.

[66] Song F, Chen P, Sun D, Wang M, Dong L, Liang D, Xu R-M, Zhu P, Li G. Cryo-EM study of the chromatin fiber reveals a double helix twisted by tetranucleosomal units. Science (2014), 344:376–380.

[67] Almer A, Rudolph H, Hinnen A, Hörz W. Removal of positioned nucleosomes from the yeast PHO5 promoter upon PHO5 induction releases additional upstream activating DNA elements. EMBO J (1986), 5:2689–2696.

[68] Richard-Foy H, Hager GL. Sequence-specific positioning of nucleosomes over the steroid-inducible MMTV promoter. EMBO J (1987), 6:2321–2328.

[69] Kornberg RD, Stryer L. Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism. Nucleic Acids Res (1988), 16:6677–6690.

[70] Hewish DR, Burgoyne LA. Chromatin sub-structure. The digestion of chromatin DNA at regularly spaced sites by a nuclear deoxyribonuclease. Biochemical and Biophysical Research Communications (1973), 52:504–510.

[71] Noll M. Subunit structure of chromatin. Nature (1974), 251:249–251.

[72] Kornberg RD, Thomas JO. Chromatin structure; oligomers of the histones. Science (1974), 184:865–868.

[73] Noll M, Kornberg RD. Action of micrococcal nuclease on chromatin and the location of histone H1. J Mol Biol (1977), 109:393–404.

[74] 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. J Mol Biol (1986),

190:619–633.

[75] Hertzberg RP, Dervan PB. Cleavage of DNA with methidiumpropyl-EDTA-iron(II): reaction conditions and product analyses. Biochemistry (1984), 23:3934–3945.

[76] Cartwright IL, Hertzberg RP, Dervan PB, Elgin SC. Cleavage of chromatin with methidiumpropyl-EDTA.iron(II). Proc Natl Acad Sci USA (1983), 80:3213–3217.

[77] Cartwright IL, Elgin SC. Analysis of chromatin structure and DNA sequence organization: use of the 1,10-phenanthroline-cuprous complex. Nucleic Acids Res (1982), 10:5835–5852.

[78] 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:e15754.

[79] Gaffney DJ, McVicker G, Pai AA, Fondufe-Mittendorf YN, Lewellen N, Michelini K, Widom J, Gilad Y, Pritchard JK. Controls of nucleosome positioning in the human genome. PLoS Genet (2012), 8:e1003036.

[80] Wal M, Pugh BF. Genome-wide mapping of nucleosome positions in yeast using high-resolution MNase ChIP-Seq. Meth Enzymol (2012), 513:233–250.

[81] Allan J, Fraser RM, Owen-Hughes T, Keszenman-Pereyra D. Micrococcal Nuclease Does Not Substantially Bias Nucleosome Mapping. J Mol Biol (2012), 417:152–164.

[82] Thoma F, Bergman LW, Simpson RT. Nuclease digestion of circular TRP1ARS1 chromatin reveals positioned nucleosomes separated by nuclease-sensitive regions. J Mol Biol (1984), 177:715–733.

[83] Shimizu M, Roth SY, Szent-Gyorgyi C, Simpson RT. Nucleosomes are positioned with base pair precision adjacent to the alpha 2 operator in Saccharomyces cerevisiae. EMBO J (1991), 10:3033–3041.

[84] Hayes JJ, Clark DJ, Wolffe AP. Histone contributions to the structure of DNA in the nucleosome. Proc Natl Acad Sci USA (1991), 88:6829–6833.

[85] Yenidunya A, Davey C, Clark D, Felsenfeld G, Allan J. Nucleosome positioning on chicken and human globin gene promoters in vitro. Novel mapping techniques. J Mol Biol (1994), 237:401–414.

[86] Fragoso G, John S, Roberts MS, Hager GL. Nucleosome positioning on the MMTV LTR results from the frequency-biased occupancy of multiple frames. Genes Dev (1995), 9:1933–1947.

[87] Costanzo G, Di Mauro E, Negri R, Pereira G, Hollenberg C. Multiple overlapping positions of nucleosomes with single in vivo rotational setting in the Hansenula polymorpha RNA polymerase II MOX promoter. J Biol Chem (1995), 270:11091–11097.

[88] Piña B, Brüggemeier U, Beato M. Nucleosome positioning modulates accessibility of regulatory proteins to the mouse mammary tumor virus promoter. Cell (1990), 60:719–731.

[89] Lowary PT, Widom J. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. J Mol Biol (1998), 276:19–42.

[90] Tuerk C, Gold L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science (1990), 249:505–510.

[91] Racki LR, Yang JG, Naber N, Partensky PD, Acevedo A, Purcell TJ, Cooke R, Cheng Y, Narlikar GJ. The chromatin remodeller ACF acts as a dimeric motor to space nucleosomes. Nature (2009), 462:1016–1021.

[92] Andrews AJ, Chen X, Zevin A, Stargell LA, Luger K. The histone chaperone Nap1 promotes nucleosome assembly by eliminating nonnucleosomal histone DNA interactions. Mol Cell (2010), 37:834–842.

[93] Vasudevan D, Chua EYD, Davey CA. Crystal structures of nucleosome core particles containing the "601" strong positioning sequence. J Mol Biol (2010), 403:1–10.

[94] 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), 442:772–778.

[95] Wade N. Scientists say they've found a code beyond genetics in DNA. The New York Times (2006).

[96] 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 Comput Biol (2008), 4:e1000216.

[97] Xi L, Fondufe-Mittendorf Y, Xia L, Flatow J, Widom J, Wang J-P. Predicting nucleosome positioning using a duration Hidden Markov Model. BMC Bioinformatics (2010), 11:346.

[98] Gabdank I, Barash D, Trifonov EN. FineStr: a web server for single-base-resolution nucleosome positioning. Bioinformatics (2010), 26:845–846.

[99] van der Heijden T, van Vugt JJFA, Logie C, van Noort J. Sequence-based prediction of single nucleosome positioning and genome-wide nucleosome occupancy. Proc Natl Acad Sci USA (2012), 109:E2514–22.

[100] Nikolaou C, Althammer S, Beato M, Guigó R. Structural constraints revealed in consistent nucleosome positions in the genome of S. cerevisiae. Epigenetics Chromatin (2010), 3:20.

[101] Peckham HE, Thurman RE, Fu Y, Stamatoyannopoulos JA, Noble WS, Struhl K, Weng Z. Nucleosome positioning signals in genomic DNA. Genome Res (2007), 17:1170–1177.

[102] 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. Nat Struct Mol Biol (2009), 16:847–852.

[103] 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. Proc Natl Acad Sci USA (2010), 107:17945–17950.

[104] Ohlsson R, Renkawitz R, Lobanenkov V. CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. Trends Genet (2001), 17:520–527.

[105] Wendt KS, Yoshida K, Itoh T, Bando M, Koch B, Schirghuber E,

Tsutsumi S, Nagae G, Ishihara K, Mishiro T, Yahata K, Imamoto F, Aburatani H, Nakao M, Imamoto N, Maeshima K, Shirahige K, Peters J-M. Cohesin mediates transcriptional insulation by CCCTC-binding factor. Nature (2008), 451:796–801.

[106] Splinter E, Heath H, Kooren J, Palstra R-J, Klous P, Grosveld F, Galjart N, de Laat W. CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. Genes Dev (2006), 20:2349–2354.

[107] Phillips JE, Corces VG. CTCF: master weaver of the genome. Cell (2009), 137:1194–1211.

[108] Fu Y, Sinha M, Peterson CL, Weng Z. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. PLoS Genet (2008), 4:e1000138.

[109] Chen H, Tian Y, Shu W, Bo X, Wang S. Comprehensive identification and annotation of cell type-specific and ubiquitous CTCF-binding sites in the human genome. PLoS ONE (2012), 7:e41374.

[110] Teif VB, Vainshtein Y, Caudron-Herger M, Mallm J-P, Marth C, Höfer T, Rippe K. Genome-wide nucleosome positioning during embryonic stem cell development. Nat Struct Mol Biol (2012), 19:1185–1192.

[111] Tilgner H, Nikolaou C, Althammer S, Sammeth M, Beato M, Valcárcel J, Guigó R. Nucleosome positioning as a determinant of exon recognition. Nat Struct Mol Biol (2009), 16:996–1001.

[112] Schwartz S, Meshorer E, Ast G. Chromatin organization marks exon-intron structure. Nat Struct Mol Biol (2009), 16:990–995.

[113] Andersson R, Enroth S, Rada-Iglesias A, Wadelius C, Komorowski J. Nucleosomes are well positioned in exons and carry characteristic histone modifications. Genome Res (2009), 19:1732–1741.

[114] Chodavarapu RK, Feng S, Bernatavichute YV, Chen P-Y, Stroud H, Yu Y, Hetzel JA, Kuo F, Kim J, Cokus SJ, Casero D, Bernal M, Huijser P, Clark AT, Krämer U, Merchant SS, Zhang X, Jacobsen SE, Pellegrini M. Relationship between nucleosome positioning and DNA methylation. Nature (2010).

[115] Valouev A, Johnson SM, Boyd SD, Smith CL, Fire AZ, Sidow A. De-

terminants of nucleosome organization in primary human cells. Nature (2011), 474:516–520.

[116] Xiao G, White D, Bargonetti J. p53 binds to a constitutively nucleosome free region of the mdm2 gene. Oncogene (1998), 16:1171–1181.

[117] Suter B, Schnappauf G, Thoma F. Poly(dA.dT) sequences exist as rigid DNA structures in nucleosome-free yeast promoters in vivo. Nucleic Acids Res (2000), 28:4083–4089.

[118] Henikoff S. Labile H3.3+H2A.Z nucleosomes mark 'nucleosome-free regions'. Nature Genetics (2009), 41:865–866.

[119] Schwarzbauer K, Bodenhofer U, Hochreiter S. Genome-wide chromatin remodeling identified at GC-rich long nucleosome-free regions. PLoS ONE (2012), 7:e47924.

[120] Ranjan A, Mizuguchi G, FitzGerald PC, Wei D, Wang F, Huang Y, Luk E, Woodcock CL, Wu C. Nucleosome-free region dominates histone acetylation in targeting SWR1 to promoters for H2A.Z replacement. Cell (2013), 154:1232–1245.

[121] Yen K, Vinayachandran V, Pugh BF. SWR-C and INO80 chromatin remodelers recognize nucleosome-free regions near +1 nucleosomes. Cell (2013), 154:1246–1256.

[122] Durán E, Djebali S, González S, Flores O, Mercader JM, Guigó R, Torrents D, Soler-López M, Orozco M. Unravelling the hidden DNA structural/physical code provides novel insights on promoter location. Nucleic Acids Res (2013), 41:7220–7230.

[123] Clapier CR, Cairns BR. The biology of chromatin remodeling complexes. Annu Rev Biochem (2009), 78:273–304.

[124] Saha A, Wittmeyer J, Cairns BR. Chromatin remodelling: the industrial revolution of DNA around histones. Nat Rev Mol Cell Biol (2006), 7:437–447.

[125] Cairns BR, Lorch Y, Li Y, Zhang M, Lacomis L, Erdjument-Bromage H, Tempst P, Du J, Laurent B, Kornberg RD. RSC, an essential, abundant chromatin-remodeling complex. Cell (1996), 87:1249–1260.

[126] Chaban Y, Ezeokonkwo C, Chung W-H, Zhang F, Kornberg RD,

Maier-Davis B, Lorch Y, Asturias FJ. Structure of a RSC-nucleosome complex and insights into chromatin remodeling. Nat Struct Mol Biol (2008), 15:1272–1277.

[127] Floer M, Wang X, Prabhu V, Berrozpe G, Narayan S, Spagna D, Alvarez D, Kendall J, Krasnitz A, Stepansky A, Hicks J, Bryant GO, Ptashne M. A RSC/nucleosome complex determines chromatin architecture and facilitates activator binding. Cell (2010), 141:407–418.

[128] Chawla A, Repa JJ, Evans RM, Mangelsdorf DJ. Nuclear receptors and lipid physiology: opening the X-files. Science (2001), 294:1866–1870.

[129] Greenstein BD. The role of hormone receptors in development and puberty. J Reprod Fertil (1978), 52:419–426.

[130] Debes JD, Tindall DJ. The role of androgens and the androgen receptor in prostate cancer. Cancer Lett (2002), 187:1–7.

[131] Foidart JM, Colin C, Denoo X, Desreux J, Béliard A, Fournier S, de Lignières B. Estradiol and progesterone regulate the proliferation of human breast epithelial cells. Fertil Steril (1998), 69:963–969.

[132] Taylor JA, Hirvonen A, Watson M, Pittman G, Mohler JL, Bell DA. Association of prostate cancer with vitamin D receptor gene polymorphism. Cancer Research (1996), 56:4108–4110.

[133] Guiochon-Mantel A, Lescop P, Christin-Maitre S, Loosfelt H, Perrot-Applanat M, Milgrom E. Nucleocytoplasmic shuttling of the progesterone receptor. EMBO J (1991), 10:3851–3859.

[134] Ham J, Thomson A, Needham M, Webb P, Parker M. Characterization of response elements for androgens, glucocorticoids and progestins in mouse mammary tumour virus. Nucleic Acids Res (1988), 16:5263–5276.

[135] Bittner JJ. Some possible effects of nursing on the mammary gland tumor incidence in mice. Science (1936), 84:162–162.

[136] Groner B, Buetti E, Diggelmann H, Hynes NE. Characterization of endogenous and exogenous mouse mammary tumor virus proviral DNA with site-specific molecular clones. J Virol (1980), 36:734–745.

[137] Ucker DS, Ross SR, Yamamoto KR. Mammary tumor virus DNA

contains sequences required for its hormone-regulated transcription. Cell (1981), 27:257–266.

[138] Buetti E, Diggelmann H. Glucocorticoid regulation of mouse mammary tumor virus: identification of a short essential DNA region. EMBO J (1983), 2:1423–1429.

[139] Cato AC, Miksicek R, Schütz G, Arnemann J, Beato M. The hormone regulatory element of mouse mammary tumour virus mediates progesterone induction. EMBO J (1986), 5:2237–2240.

[140] Cato AC, Henderson D, Ponta H. The hormone response element of the mouse mammary tumour virus DNA mediates the progestin and androgen induction of transcription in the proviral long terminal repeat region. EMBO J (1987), 6:363–368.

[141] Chávez S, Candau R, Truss M, Beato M. Constitutive repression and nuclear factor I-dependent hormone activation of the mouse mammary tumor virus promoter in Saccharomyces cerevisiae. Mol Cell Biol (1995), 15:6987–6998.

[142] Venditti P, Di Croce L, Kauer M, Blank T, Becker PB, Beato M. Assembly of MMTV promoter minichromosomes with positioned nucleosomes precludes NF1 access but not restriction enzyme cleavage. Nucleic Acids Res (1998), 26:3657–3666.

[143] Horwitz KB, Mockus MB, Lessey BA. Variant T47D human breast cancer cells with high progesterone-receptor levels despite estrogen and antiestrogen resistance. Cell (1982), 28:633–642.

[144] Migliaccio A, Piccolo D, Castoria G, Di Domenico M, Bilancio A, Lombardi M, Gong W, Beato M, Auricchio F. Activation of the Src/p21ras/Erk pathway by progesterone receptor via cross-talk with estrogen receptor. EMBO J (1998), 17:2008–2018.

[145] Proietti C, Salatino M, Rosemblit C, Carnevale R, Pecci A, Kornblihtt AR, Molinolo AA, Frahm I, Charreau EH, Schillaci R, Elizalde PV. Progestins induce transcriptional activation of signal transducer and activator of transcription 3 (Stat3) via a Jak- and Src-dependent mechanism in breast cancer cells. Mol Cell Biol (2005), 25:4826–4840.

[146] Ballaré C, Castellano G, Gaveglia L, Althammer S, González-Vallinas

J, Eyras E, Le Dily F, Zaurin R, Soronellas D, Vicent GP, Beato M. Nucleosome-Driven Transcription Factor Binding and Gene Regulation. Mol Cell (2013).

[147] Kuo MH, Allis CD. In vivo cross-linking and immunoprecipitation for studying dynamic Protein:DNA associations in a chromatin environment. Methods (1999), 19:425–433.

[148] Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA. Genome-wide location and function of DNA binding proteins. Science (2000), 290:2306–2309.

[149] Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim T-K, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature (2007), 448:553–560.

[150] Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science (2007), 316:1497–1502.

[151] ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol (2011), 9:e1001046.

[152] Einhauer A, Jungbauer A. The FLAG peptide, a versatile fusion tag for the purification of recombinant proteins. J Biochem Biophys Methods (2001), 49:455–465.

[153] Field J, Nikawa J, Broek D, MacDonald B, Rodgers L, Wilson IA, Lerner RA, Wigler M. Purification of a RAS-responsive adenylyl cyclase complex from Saccharomyces cerevisiae by use of an epitope addition method. Mol Cell Biol (1988), 8:2159–2165.

[154] Hilpert K, Hansen G, Wessner H, Küttner G, Welfle K, Seifert M, Höhne W. Anti-c-myc antibody 9E10: epitope key positions and variability characterized using peptide spot synthesis on cellulose. Protein Eng (2001), 14:803–806.

[155] van Steensel B, Henikoff S. Identification of in vivo DNA targets of

chromatin proteins using tethered dam methyltransferase. Nat Biotechnol (2000), 18:424–428.

[156] van Steensel B, Delrow J, Henikoff S. Chromatin profiling using targeted DNA adenine methyltransferase. Nature Genetics (2001), 27:304–308.

[157] Licatalosi DD, Mele A, Fak JJ, Ule J, Kayikci M, Chi SW, Clark TA, Schweitzer AC, Blume JE, Wang X, Darnell JC, Darnell RB. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature (2008), 456:464–469.

[158] Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E, Chang HY. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell (2007), 129:1311–1323.

[159] Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Rivea Morales D, Thomas K, Presser A, Bernstein BE, van Oudenaarden A, Regev A, Lander ES, Rinn JL. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci USA (2009), 106:11667–11672.

[160] Hendrickson DG, Hogan DJ, McCullough HL, Myers JW, Herschlag D, Ferrell JE, Brown PO. Concordant regulation of translation and mRNA abundance for hundreds of targets of a human microRNA. PLoS Biol (2009), 7:e1000238.

[161] Hendrickson DG, Hogan DJ, Herschlag D, Ferrell JE, Brown PO. Systematic identification of mRNAs recruited to argonaute 2 by specific microRNAs and corresponding changes in transcript abundance. PLoS ONE (2008), 3:e2126.

[162] Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. (2011), 44:667–678.

[163] Wu C. The 5' ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I. Nature (1980), 286:854–860.

[164] Keene MA, Corces V, Lowenhaupt K, Elgin SC. DNase I hypersensitive sites in Drosophila chromatin occur at the 5' ends of regions of transcription. Proc Natl Acad Sci USA (1981), 78:143–146.

[165] Bryan PN, Folk WR. Enhancer sequences responsible for DNase I hypersensitivity in polyomavirus chromatin. Mol Cell Biol (1986), 6:2249–2252.

[166] Song L, Crawford GE. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harb Protoc (2010), 2010:pdb.prot5384–pdb.prot5384.

[167] High-resolution mapping and characterization of open chromatin across the genome. (2008), 132:311–322.

[168] Crawford GE, Holt IE, Whittle J, Webb BD, Tai D, Davis S, Margulies EH, Chen Y, Bernat JA, Ginsburg D, Zhou D, Luo S, Vasicek TJ, Daly MJ, Wolfsberg TG, Collins FS. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome Res (2006), 16:123–131.

[169] Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, Garg K, John S, Sandstrom R, Bates D, Boatman L, Canfield TK, Diegel M, Dunn D, Ebersol AK, Frum T, Giste E, Johnson AK, Johnson EM, Kutyavin T, Lajoie B, Lee B-K, Lee K, London D, Lotakis D, Neph S, Neri F, Nguyen ED, Qu H, Reynolds AP, Roach V, Safi A, Sanchez ME, Sanyal A, Shafer A, Simon JM, Song L, Vong S, Weaver M, Yan Y, Zhang Z, Zhang Z, Lenhard B, Tewari M, Dorschner MO, Hansen RS, Navas PA, Stamatoyannopoulos G, Iyer VR, Lieb JD, Sunyaev SR, Akey JM, Sabo PJ, Kaul R, Furey TS, Dekker J, Crawford GE, Stamatoyannopoulos JA. The accessible chromatin landscape of the human genome. Nature (2012), 489:75–82.

[170] Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res (2007), 17:877–885.

[171] Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee B-K, Sheffield NC, Gräf S, Huss M, Keefe D, Liu Z, London D, McDaniell RM, Shibata Y, Showers KA, Simon JM, Vales T, Wang T, Winter D, Zhang

Z, Clarke ND, Birney E, Iyer VR, Crawford GE, Lieb JD, Furey TS. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Res (2011), 21:1757–1767.

[172] Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. Science (2002), 295:1306–1311.

[173] Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, Rubio ED, Krumm A, Lamb J, Nusbaum C, Green RD, Dekker J. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res (2006), 16:1299–1309.

[174] Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science (2009), 326:289–293.

[175] de Wit E, de Laat W. A decade of 3C technologies: insights into nuclear organization. Genes Dev (2012), 26:11–24.

[176] Lipman DJ, Pearson WR. Rapid and sensitive protein similarity searches. Science (1985), 227:1435–1441.

[177] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol (1990), 215:403–410.

[178] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics (2009), 25:2078–2079.

[179] Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res (2013):gkt1030.

[180] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res (2002), 30:207–210.

[181] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky

M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res (2013), 41(Database issue):D991–5.

[182] Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, Oezcimen A, Rocca-Serra P, Sansone S-A. ArrayExpress--a public repository for microarray gene expression data at the EBI. Nucleic Acids Res (2003), 31:68–71.

[183] Rustici G, Kolesnikov N, Brandizi M, Burdett T, Dylag M, Emam I, Farne A, Hastings E, Ison J, Keays M, Kurbatova N, Malone J, Mani R, Mupo A, Pedro Pereira R, Pilicheva E, Rung J, Sharma A, Tang YA, Ternent T, Tikhonov A, Welter D, Williams E, Brazma A, Parkinson H, Sarkans U. ArrayExpress update--trends in database growth and links to data analysis tools. Nucleic Acids Res (2013), 41(Database issue):D987–90.

[184] Sherlock G, Hernandez-Boussard T, Kasarskis A, Binkley G, Matese JC, Dwight SS, Kaloper M, Weng S, Jin H, Ball CA, Eisen MB, Spellman PT, Brown PO, Botstein D, Cherry JM. The Stanford Microarray Database. Nucleic Acids Res (2001), 29:152–155.

[185] Mei R, Galipeau PC, Prass C, Berno A, Ghandour G, Patil N, Wolff RK, Chee MS, Reid BJ, Lockhart DJ. Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays. Genome Res (2000), 10:1126–1137.

[186] Kennedy GC, Matsuzaki H, Dong S, Liu W-M, Huang J, Liu G, Su X, Cao M, Chen W, Zhang J, Liu W, Yang G, Di X, Ryder T, He Z, Surti U, Phillips MS, Boyce-Jacino MT, Fodor SPA, Jones KW. Large-scale genotyping of complex DNA. Nat Biotechnol (2003), 21:1233–1237.

[187] Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, Garcia F, Haden K, Li J, Shaw CA, Belmont J, Cheung SW, Shen RM, Barker DL, Gunderson KL. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. Genome Res (2006), 16:1136–1148.

[188] Clark TA, Schweitzer AC, Chen TX, Staples MK, Lu G, Wang H, Williams A, Blume JE. Discovery of tissue-specific exons using comprehensive human exon microarrays. Genome Biol (2007), 8:R64.

[189] Robinson MD, Speed TP. Differential splicing using whole-transcript microarrays. BMC Bioinformatics (2009), 10:156.

[190] Skotheim RI, Thomassen GOS, Eken M, Lind GE, Micci F, Ribeiro FR, Cerveira N, Teixeira MR, Heim S, Rognes T, Lothe RA. A universal assay for detection of oncogenic fusion transcripts by oligo microarray analysis. Mol Cancer (2009), 8:5.

[191] Schuster SC. Next-generation sequencing transforms today's biology. Nat Methods (2008), 5:16–18.

[192] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet (2009), 10:57–63.

[193] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods (2008), 5:621–628.

[194] Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol (2010), 28:511–515.

[195] Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. Model-based analysis of ChIP-Seq (MACS). Genome Biol (2008), 9:R137.

[196] Fejes AP, Robertson G, Bilenky M, Varhol R, Bainbridge M, Jones SJM. FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. Bioinformatics (2008), 24:1729–1730.

[197] Boyle AP, Guinney J, Crawford GE, Furey TS. F-Seq: a feature density estimator for high-throughput sequence tags. Bioinformatics (2008), 24:2537–2538.

[198] Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. Simple combinations of lineage-de-

termining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell (2010), 38:576–589.

[199] Althammer S, González-Vallinas J, Ballaré C, Beato M, Eyras E. Pyicos: a versatile toolkit for the analysis of high-throughput sequencing data. Bioinformatics (2011), 27:3333–3340.

[200] Raney BJ, Dreszer TR, Barber GP, Clawson H, Fujita PA, Wang T, Nguyen N, Paten B, Zweig AS, Karolchik D, Kent WJ. Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. Bioinformatics (2013).

[201] Abbott A. Europe to map the human epigenome. Nature (2011):518–518.

[202] Zhou X, Wang T. Using the Wash U Epigenome Browser to examine genome-wide sequencing data. Curr Protoc Bioinformatics (2012), Chapter 10:Unit10.10–10.10.14.

[203] Chadwick LH. The NIH Roadmap Epigenomics Program data resource. Epigenomics (2012), 4:317–324.

[204] Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A. Galaxy: a platform for interactive large-scale genome analysis. Genome Res (2005), 15:1451–1455.

[205] Jansen A, Verstrepen KJ. Nucleosome positioning in Saccharomyces cerevisiae. Microbiol Mol Biol Rev (2011), 75:301–320.

[206] Flemming W. Zellsubstanz, Kern Und Zelltheilung. Leipzig, F. C. W. Vogel; (1882).

[207] Luger K, Dechassa ML, Tremethick DJ. New insights into nucleosome and chromatin structure: an ordered state or a disordered affair? Nature reviews Molecular cell ... (2012).

[208] Corona DF, Längst G, Clapier CR, Bonte EJ, Ferrari S, Tamkun JW, Becker PB. ISWI is an ATP-dependent nucleosome remodeling factor. Mol Cell (1999), 3:239–245.

[209] Kundaje A, Kyriazopoulou-Panagiotopoulou S, Libbrecht M, Smith CL, Raha D, Winters EE, Johnson SM, Snyder M, Batzoglou S, Sidow A. Ubiquitous heterogeneity and asymmetry of the chromatin envi-

ronment at regulatory elements. Genome Res (2012), 22:1735–1747.
[210] Ghorbani M, Mohammad-Rafiee F. Geometrical correlations in the nucleosomal DNA conformation and the role of the covalent bonds rigidity. Nucleic Acids Res (2011), 39:1220–1230.
[211] Yun M, Wu J, Workman JL, Li B. Readers of histone modifications. Cell Res (2011), 21:564–578.
[212] van Steensel B. Mapping of genetic and epigenetic regulatory networks using microarrays. Nature Genetics (2005), 37 Suppl:S18–24.
[213] Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, van Steensel B, de Laat W. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nature Genetics (2006), 38:1348–1354.
[214] Splinter E, de Wit E, Nora EP, Klous P, van de Werken HJG, Zhu Y, Kaaij LJT, van Ijcken W, Gribnau J, Heard E, de Laat W. The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. Genes Dev (2011), 25:1371–1383.
[215] Horike S-I, Cai S, Miyano M, Cheng J-F, Kohwi-Shigematsu T. Loss of silent-chromatin looping and impaired imprinting of DLX5 in Rett syndrome. Nature Genetics (2005), 37:31–40.
[216] Fullwood MJ, Wei C-L, Liu ET, Ruan Y. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. Genome Res (2009), 19:521–532.
[217] Siepel A, Pollard KS, Haussler D. New Methods for Detecting Lineage-Specific Selection. RECOMB (2006):190–205.
[218] Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AFA, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, Haussler D, Miller W. Aligning multiple genomic sequences with the threaded block*set ali*gner. Genome Res (2004), 14:708–715.
[219] Raney BJ, Cline MS, Rosenbloom KR, Dreszer TR, Learned K, Barber GP, Meyer LR, Sloan CA, Malladi VS, Roskin KM, Suh BB, Hinrichs AS, Clawson H, Zweig AS, Kirkup V, Fujita PA, Rhead B, Smith KE, Pohl A, Kuhn RM, Karolchik D, Haussler D, Kent WJ. ENCODE whole-genome data in the UCSC genome browser (2011

update). Nucleic Acids Res (2011), 39(Database issue):D871–5.

[220] Truss M, Bartsch J, Schelbert A, Haché RJ, Beato M. Hormone induces binding of receptors and transcription factors to a rearranged nucleosome on the MMTV promoter in vivo. EMBO J (1995), 14:1737–1751.

[221] Di Croce L, Koop R, Venditti P, Westphal HM, Nightingale KP, Corona DF, Becker PB, Beato M. Two-step synergism between the progesterone receptor and the DNA-binding domain of nuclear factor 1 on MMTV minichromosomes. Mol Cell (1999), 4:45–54.

[222] Vicent GP, Nacht AS, Smith CL, Peterson CL, Dimitrov S, Beato M. DNA instructed displacement of histones H2A and H2B at an inducible promoter. Mol Cell (2004), 16:439–452.

[223] Vicent GP, Nacht AS, Font-Mateu J, Castellano G, Gaveglia L, Ballaré C, Beato M. Four enzymes cooperate to displace histone H1 during the first minute of hormonal gene activation. Genes Dev (2011), 25:845–862.

[224] Vicent GP, Nacht AS, Zaurin R, Ballare C, Clausell J, Beato M. Minireview: Role of Kinases and Chromatin Remodeling in Progesterone Signaling to Chromatin. Molecular Endocrinology (2010), 24:2088–2098.

[225] Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics (2010), 26:589–595.

[226] Pickrell JK, Gaffney DJ, Gilad Y, Pritchard JK. False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions. Bioinformatics (2011), 27:2144–2146.

[227] n-Lagos MR, Di Cantogno LV, Marchi C, Rangel N, Payan-Gomez C, Gugliotta P, Botta C, Bussolati G, rez-Clavijo SRR, Pasini B, Sapino A. Differences and homologies of chromosomalalterations within and between breast cancer celllines: a clustering analysis. Molecular Cytogenetics (2014), 7:1–14.

[228] Derrien T, Estellé J, Marco-Sola S, Knowles DG, Raineri E, Guigó R, Ribeca P. Fast computation and applications of genome mappability.

PLoS ONE (2012), 7:e30377.

[229] Pohl A, Beato M. bwtool: A tool for bigWig files. Bioinformatics (2014):btu056.

[230] Xi Y, Yao J, Chen R, Li W, He X. Nucleosome fragility reveals novel functional states of chromatin and poises genes for activation. Genome Res (2011), 21:718–724.

[231] Jin C, Felsenfeld G. Nucleosome stability mediated by histone variants H3.3 and H2A.Z. Genes Dev (2007), 21:1519–1529.

[232] H3.3/H2A.Z double variant-containing nucleosomes mark "nucleosome-free regions" of active promoters and other regulatory regions. (2009), 41:941–945.

[233] Konev AY, Tribus M, Park SY, Podhraski V, Lim CY, Emelyanov AV, Vershilova E, Pirrotta V, Kadonaga JT, Lusser A, Fyodorov DV. CHD1 motor protein is required for deposition of histone variant H3.3 into chromatin in vivo. Science (2007), 317:1087–1090.

[234] Afek A, Lukatsky DB. Genome-wide organization of eukaryotic preinitiation complex is influenced by nonconsensus protein-DNA binding. Biophys J (2013), 104:1107–1115.

[235] Ioshikhes I, Trifonov EN, Zhang MQ. Periodical distribution of transcription factor sites in promoter regions and connection with chromatin structure. Proc Natl Acad Sci USA (1999), 96:2891–2895.

[236] Goñi JR, Pérez A, Torrents D, Orozco M. Determining promoter location based on DNA structure first-principles calculations. Genome Biol (2007), 8:R263.

[237] Greenbaum JA, Pang B, Tullius TD. Construction of a genome-scale structural map at single-nucleotide resolution. Genome Res (2007), 17:947–953.

[238] Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods (2013), 10:1213–1218.

[239] Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. DAVID: Database for Annotation, Visualization, and Inte-

grated Discovery. Genome Biol (2003), 4:P3.

[240] 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), 458:362–366.

[241] Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res (2005), 15:1034–1050.

[242] Wright RHG, Castellano G, Bonet J, Le Dily F, Font-Mateu J, Ballaré C, Nacht AS, Soronellas D, Oliva B, Beato M. CDK2-dependent activation of PARP-1 is required for hormonal gene regulation in breast cancer cells. Genes Dev (2012), 26:1972–1983.

[243] Cao K, Lailler N, Zhang Y, Kumar A, Uppal K, Liu Z, Lee EK, Wu H, Medrzycki M, Pan C, Ho P-Y, Cooper GP, Dong X, Bock C, Bouhassira EE, Fan Y. High-resolution mapping of h1 linker histone variants in embryonic stem cells. PLoS Genet (2013), 9:e1003417.

[244] Millan-Arino L, Islam ABMMK, Izquierdo-Bouldstridge A, Mayor R, Terme JM, Luque N, Sancho M, Lopez-Bigas N, Jordan A. Mapping of six somatic linker histone H1 variants in human breast cancer cells uncovers specific features of H1.2. Nucleic Acids Res (2014):gku079.

[245] Sancho M. Role of linker Histone H1 variants in cell proliferation, Chromatin Structure and Gene expression in breast cancer cells. UPF PhD Theses (2008):1–160.

[246] Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its new supplement TREMBL. Nucleic Acids Res (1996), 24:21–25.

[247] Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita PA, Diekhans M, Smith KE, Rosenbloom KR, Raney BJ, Pohl A, Pheasant M, Meyer LR, Learned K, Hsu F, Hillman-Jackson J, Harte RA, Giardine B, Dreszer TR, Clawson H, Barber GP, Haussler D, Kent WJ. The UCSC Genome Browser database: update 2010. Nucleic Acids

Res (2010), 38(Database issue):D613–9.

[248] Le Dily F, Baù D, Pohl A, Vicent GP, Soronellas D, Castellano G, Serra F, Wright RH, Ballaré C, Filion GJ, Marti-Renom M, Beato M. Hormone elicits structural reorganization of distinct topological domains in the breast cancer genome. Genes Dev (2014).

[249] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics (2010), 26:841–842.

[250] Chadee DN, Allis CD, Wright JA, Davie JR. Histone H1b phosphorylation is dependent upon ongoing transcription and replication in normal and ras-transformed mouse fibroblasts. J Biol Chem (1997), 272:8113–8116.

[251] Bolzer A, Kreth G, Solovei I, Koehler D, Saracoglu K, Fauth C, Müller S, Eils R, Cremer C, Speicher MR, Cremer T. Three-Dimensional Maps of All Chromosomes in Human Male Fibroblast Nuclei and Prometaphase Rosettes. PLoS Biol (2005), 3:e157.

[252] Parada LA, Misteli T. Chromosome positioning in the interphase nucleus. Trends Cell Biol (2002), 12:1–8.

[253] Xing H, Mo Y, Liao W, Zhang MQ. Genome-wide localization of protein-DNA binding and histone modification by a Bayesian change-point method with ChIP-seq data. PLoS Comput Biol (2012), 8:e1002613.

[254] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res (2002), 12:996–1006.

[255] Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, García-Girón C, Gordon L, Hourlier T, Hunt S, Juettemann T, Kähäri AK, Keenan S, Komorowska M, Kulesha E, Longden I, Maurel T, McLaren WM, Muffato M, Nag R, Overduin B, Pignatelli M, Pritchard B, Pritchard E, Riat HS, Ritchie GRS, Ruffier M, Schuster M, Sheppard D, Sobral D, Taylor K, Thormann A, Trevanion S, White S, Wilder SP, Aken BL, Birney E, Cunningham F, Dunham I, Harrow J, Herrero J, Hubbard TJP, Johnson N, Kinsella R, Parker A, Spudich G, Yates

A, Zadissa A, Searle SMJ. Ensembl 2013. Nucleic Acids Res (2013), 41(Database issue):D48–55.

[256] Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics (2010), 26:2204–2207.

[257] Li H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics (2011), 27:718–719.

[258] Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods (2010), 7:1009–1015.

[259] Kuhn RM, Haussler D, Kent WJ. The UCSC genome browser and associated tools. Brief Bioinformatics (2013), 14:144–161.

[260] Shieh J, Keogh E. iSAX: disk-aware mining and indexing of massive time series datasets. Data Mining and Knowledge Discovery (2009), 19:24–57.

[261] Tevethia MJ, Ozer HL. SV40-mediated immortalization. Methods Mol Biol (2001), 165:185–199.

[262] Schones DE, Cui K, Cuddapah S, Roh T-Y, Barski A, Wang Z, Wei G, Zhao K. Dynamic regulation of nucleosome positioning in the human genome. Cell (2008), 132:887–898.

[263] ENCODE Project Consortium, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. Nature (2012), 489:57–74.

[264] Nikolaou C, Zaurin R, Althammer S, Rué P, Guigó R, Beato M. Symmetry of DNA curvature contributes to the positioning of key nucleosomes. (2010):1–35.

[265] Rhee HS, Pugh BF. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell (2011), 147:1408–1419.

[266] Déjardin J, Kingston RE. Purification of proteins associated with specific genomic Loci. Cell (2009), 136:175–186.

[267] Byrum SD, Raman A, Taverna SD, Tackett AJ. ChAP-MS: a method for identification of proteins and histone posttranslational modifica-

tions at a single genomic locus. CellReports (2012), 2:198–205.

[268] Giepmans BNG, Adams SR, Ellisman MH, Tsien RY. The fluorescent toolbox for assessing protein location and function. Science (2006), 312:217–224.

[269] Poser I, Sarov M, Hutchins JRA, Hériché J-K, Toyoda Y, Pozniakovsky A, Weigl D, Nitzsche A, Hegemann B, Bird AW, Pelletier L, Kittler R, Hua S, Naumann R, Augsburg M, Sykora MM, Hofemeister H, Zhang Y, Nasmyth K, White KP, Dietzel S, Mechtler K, Durbin R, Stewart AF, Peters J-M, Buchholz F, Hyman AA. BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals. Nat Methods (2008), 5:409–415.

[270] Zhang F, Cong L, Lodato S, Kosuri S, Church GM, Arlotta P. Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. Nat Biotechnol (2011), 29:149–153.

[271] Smith J, Bibikova M, Whitby FG, Reddy AR, Chandrasegaran S, Carroll D. Requirements for double-strand cleavage by chimeric restriction enzymes with zinc finger DNA-recognition domains. Nucleic Acids Res (2000), 28:3361–3369.

[272] Bibikova M, Carroll D, Segal DJ, Trautman JK, Smith J, Kim YG, Chandrasegaran S. Stimulation of homologous recombination through targeted cleavage by chimeric nucleases. Mol Cell Biol (2001), 21:289–297.

[273] Pennisi E. The CRISPR craze. Science (2013):833–836.

[274] Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA, Zhang F. Multiplex genome engineering using CRISPR/Cas systems. Science (2013), 339:819–823.

[275] DiCarlo JE, Norville JE, Mali P, Rios X, Aach J, Church GM. Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. Nucleic Acids Res (2013), 41:4336–4343.

[276] Hwang WY, Fu Y, Reyon D, Maeder ML, Tsai SQ, Sander JD, Peterson RT, Yeh J-RJ, Joung JK. Efficient genome editing in zebrafish using a CRISPR-Cas system. Nat Biotechnol (2013), 31:227–229.

[277] Thomson JA, Itskovitz-Eldor J, Shapiro SS, Waknitz MA, Swiergiel JJ,

Marshall VS, Jones JM. Embryonic stem cell lines derived from human blastocysts. Science (1998), 282:1145–1147.

[278] Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, Rodriguez JM, Ezkurdia I, van Baren J, Brent M, Haussler D, Kellis M, Valencia A, Reymond A, Gerstein M, Guigó R, Hubbard TJ. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res (2012), 22:1760–1774.

Abreviaturas

3C	Chromosome conformation capture.
ATAC-seq	Sequencing experiment from the result of an Assay for Transposase-Accessible Chromatin.
ATP	Adenosine-triphosphate: a molecule metabolized for many cellular processes.
BAC	Bacterial Artificial Chromosome, a large DNA construct.
BAF	Brg1/Brm associated factors.
BAM	Binary SAM.
BLAST	Basic Local Alignment Search Tool.
bp	Basepair(s).
BWA	Burrows-Wheeler Aligner, for mapping short sequence reads to the genome.
ChIP	Chromatin immunoprecipitation.
ChIP-seq	Sequencing experiment of DNA resulting from a ChIP.
CpG	C-phosphate-G: a cytosine next to a guanine separated by a phosphate in linear sequence; to distinguish from CG base-pairing.
CMV	Cytomegalovirus promoter is a constitutive mammalian promoter for driving the expression of transgenes.
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats, but also generally denotes the genomic editing method using these DNA sequences and Cas proteins.
CTCF	CCCTC-binding factor or 11-zinc finger protein.
CTD	Carboxyl terminal domain of a protein: the COOH at the end of an amino acid chain.
DAPI	4',6-diamidino-2-phenylindole is a fluorescent stain that binds strongly DNA.
DamID	DNA adenine methyltransferase identification.
DSB	Double-stranded break (in DNA).
DNase	Deoxyribonuclease

DNase-seq	Sequencing experiment of the regions of DNA assayed to be DNase sensitive or hypersensitive.
ENCODE	Encyclopedia of DNA Elements: an international consortium for the annotation of regulatory elements in the human genome.
FACS	Fluorescence-activated Cell Sorting; a method using flow cytometry.
FDR	False discovery rate.
FISH/M-FISH	Fluorescence *in situ* hybridization/multicolor FISH.
FTP	File Transfer Protocol: an archaic part of the internet still in used, mainly for the repositories of large data files.
G0/G1 /S/G2/M	Cell cycle phases in a cell. Gap 0: Resting phase; Gap1: growth and synthesis checkpoint; S: synthesis (DNA replication); G2: growth and mitosis checkpoint; M: mitosis, cell division.
Gb/Gbase	Gigabasepair, or $1x10^9$ basepairs of DNA.
GC	Guanine/cystosine base-paring in DNA.
GEO	Gene Expression Omnibus database.
GNU	A free software, massively collaborative software project associated with Linux and UNIX that advocates free software and also provides generic licenses for people unfamiliar with copyright law.
GFP	Green fluorescence protein.
HA	Human influenza hemagglutinin surface glycoprotein used as an epitope tag.
HAT	Histone acetyltransferase.
HeLa	A human immortal cell line derived from cervical cancer cells taken from Henrietta Lacks in 1951.
hESC/H1 hESC	Human embryonic stem cells. H1 hESC refers to the cell line established at the University of Wisconsin – Madison in 1998.
HDAC	Histone deacetylase.
HDR	Highly-duplicated region.

hg19	Human genome reference, UCSC version 19, corresponding to the Feb 2009 release (GRCh37) of the Genome Reference Consortium.
Hi-C	High-throughput 3C.
HP1	Heterochromatin protein 1, also known as Chromobox Homolog or CBX.
HOMER	NGS analysis toolkit from the Salk Institute.
HRE	Hormone responsive element.
HS	Hypersensitive.
HTS	High-throughput sequencing, also known as NGS or deep sequencing.
HTTP	Hypertext Transfer Protocol: the part of the internet used by the world-wide web.
ISWI	Imitation SWI: *D. melanogaster* ATPase the ISWI family of chromatin remodelers is named for, which include NURF, CHRAC, and ACF.
kb/Kbase	Kilobasepair, or 1,000 basepairs of DNA.
LC-MS/MS	Liquid chromatography-tandem mass spectrometry.
MACS	Model-based Analysis of ChIP-seq.
Mb/Mbase	Megabasepair, or $1x10^6$ basepairs of DNA.
MNase	Microccocal nuclease.
MNase-seq	Sequencing experiment of DNA resulting from an MNase digestion.
MMTV	Sequencing experiment of DNA resulting from an MNase digestion.
NDR	Nucleosome-depleted region.
NFR	Nucleosome-free region.
NGS	Next-generation sequencing, also known as HTS, or deep sequencing.
NTD	Amine terminal domain of a protein: the NH2 at the beginning of an amino acid chain.
PBS	Phosphate buffered saline.
PCR	Polymerase chain reaction.

PRC1/PRC2	Polycomb repressive complexes 1 and 2.
PTM	Post-translational modification.
RNA-seq	Sequencing experiment of RNA-derived cDNA taken from the RNA from cells or subcellular compartments.
RPKM	Reads per kilobase per million, a rate defining a quantification of gene expression based on RNA-seq data.
SAM	Sequence Alignment Map.
SAX	Symbolic Aggregate approXimation: a discretization method i.e. a method that change numerically-based data into text.
SELEX	Systematic evolution of ligands by exponential enrichment.
siRNA	Small Interfering RNA.
SNP	Single-nucleotide polymorphism.
SV40	Simian vacuolating virus 40.
SWI/SNF	SWItch/Sucrose NonFermentable chromatin remodeler found in eukaryotes and prokaryotes.
T47D	A human ductal breast epithelial tumor cell line.
TAD	Topologically associating domain
Tb/Tbase	Terabasepair, or $1x10^{12}$ basepairs of DNA.
TSS	Transcription start site.
TTS	Transcription termination site.
UCSC	University of California – Santa Cruz.
URL	Unique Record Locator: an address on the internet for a web page or other type of server site (e.g. FTP site).
WIG	Wiggle: a text-based file format for continuous genomic signal data.

Índex de figures (Nucleosome Zero)

Índex de figures (Histone H1)