Aus dem Adolf-Butenandt-Institut
der Ludwig-Maximilians-Universität München
Lehrstuhl Molekularbiologie im Biomedizinischen Centrum (BMC)
Vorstand: Prof. Dr. rer. nat. Peter B. Becker

Cell-free Genomics Reveal Intrinsic, Cooperative and Competitive Interactions of Chromatin Binding Proteins

Dissertation
zum Erwerb des Doktorgrades der Naturwissenschaften
an der Medizinischen Fakultät der
Ludwig-Maximilians-Universität München

vorgelegt von
Nikolas Christof Maximilian Eggers

aus
Bad Soden am Taunus

2021

Mit Genehmigung der Medizinischen Fakultät der
Universität München

Betreuer:	Prof. Dr. rer. nat. Peter B. Becker
Zweitgutachter:	Prof. Dr. rer. nat. Stefan Stricker
Dekan:	Prof. Dr. med. Thomas Gudermann
Tag der mündlichen Prüfung:	25.03.2022

Eidesstattliche Versicherung

Eggers, Nikolas

Ich erkläre hiermit an Eides statt,
dass ich die vorliegende Dissertation mit dem Thema

Cell-free Genomics Reveal Intrinsic, Cooperative and Competitive Interactions of Chromatin Binding Proteins

selbständig verfasst, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

Ich erkläre des Weiteren, dass die hier vorgelegte Dissertation nicht in gleicher oder in ähnlicher Form bei einer anderen Stelle zur Erlangung eines akademischen Grades eingereicht wurde.

München, 31.03.2022 , Nikolas Eggers

Ort, Datum, Unterschrift

Sometimes science is more art than science.

Für meine Eltern,

Ohne deren Unterstützung diese Arbeit nicht möglich gewesen wäre.

Preface

Part of this dissertation has been published in a research article, titled “Cell-free genomics reveal intrinsic, cooperative and competitive determinants of chromatin interactions”, in Nucleic Acid Research with PMID 34181732. This includes most of the data presented in the results. I performed all the experiments and all the bioinformatic analyses.

Table of Contents

Summary

In Drosophila the transcriptional output of the single male X chromosome must be compensated to match the levels presented by the two X chromosomes in females. This process is essential and disturbances lead to male specific lethality. To this end, the Dosage Compensation Complex, consisting of 5 protein subunits [MSL1, (Male-Specific-Lethal 1), MSL2, MSL3, MOF (Males-Absent-on-the-First) and MLE (Maleless)] and a long noncoding RNA [roX (RNA on the X)], binds to high affinity sites (HAS) on the X chromosome through its DNA binding subunit MSL2. From here, the complex distributes to nearby active genes and acetylates H4K16 of their nucleosomes. This ultimately leads to a roughly 2-fold increase in expression of all X chromosomal genes. This process, therefore, depends on precise recognition of the proper binding sites, as demonstrated by MSL2 in vivo, where all binding is to the X chromosome. To this end, and similar to other transcription factors, MSL2 needs to distinguish the few functional binding sites from a large pool of seemingly similar but essentially nonfunctional binding sites. However, in in vitro assays, where MSL2 is allowed to freely select its binding sites from genomic DNA, the enrichment of binding sites to the X is only about 50%, while MSL2 still binds the same GA-rich binding motif found in vivo.

This work aims to describe which factors are missing to achieve faithful X-chromosomal recognition in vitro, to better understand the processes behind Dosage Compensation Complex targeting in particular and transcription factor binding in general. The results of this work are divided into three major parts. Firstly, it describes “cell-free genomics”. Using a preblastoderm extract of Drosophila embryos (DREX) chromatin is reconstituted on genomic Drosophila DNA in vitro, which can be used as a physiological substrate for transcription factor and nucleosome remodelers to interact with. Secondly, MSL2 and a known MSL2 interacting protein, CLAMP (chromatin linked adaptor for MSL proteins), are probed regarding interactions with each other and the DREX-assembled chromatin substrate in vitro. These proteins have been well characterized in vitro and provide a reliable control to test DREX-assembled chromatin. Cell-free genomics allow to directly control the concentration of relevant factors, which gives novel insights into the establishment of X-chromosomal targeting and transcriptions factor cooperation and competition. Here, MSL2 is uniquely suited as a model protein to distinguish between productive and non-productive DNA binding, as all its functional binding sites are located on the X chromosome by definition. Lastly, bioinformatical analyses are performed revealing how complex shape features influence how transcription factors distinguish their targets from a pool of seemingly similar but nonfunctional binding sites. For MSL2 targeting, the data suggest that certain shape features at functional sites serve not as a positive recognition signal for MSL2 itself, but rather as a negative selection signal against another abundant GA-binding protein, GAF, which would otherwise outcompete MSL2. This negative selection demonstrates a novel concept of broader applicability for how the binding profile of a given transcription factor is sculpted by another unrelated factor through a complex sequence and shape signature.

Zusammenfassung

Bei Drosophila muss die Transkriptionsleistung des einzelnen X-Chromosoms in Männchen kompensiert werden, um dem Level der beiden X-Chromosomen bei den Weibchen zu entsprechen. Dieser Vorgang ist überlebensnotwendig und falls er gestört wird führt dies in Männchen zum Tod. Dazu bindet der Dosage Compensation Complex (DCC), bestehend aus 5 Proteinuntereinheiten [MSL1, (Male-Specific-Lethal 1), MSL2, MSL3, MOF (Males-Absent-on-the-First) und MLE (Maleless)] sowie einer langen nichtkodierenden RNA [roX (RNA auf dem X)], durch seine DNA-bindende Untereinheit MSL2 an Hochaffinitätsstellen (HAS) auf dem X-Chromosom. Von hier aus erreicht der Komplex benachbarte aktive Gene und acetyliert das Lysin an Position 16 in Histon 4 (H4K16) ihrer Nukleosomen. Dies führt letztendlich zu einer etwa 2-fachen Erhöhung der Expression X-chromosomaler Gene. Dieser Prozess hängt daher von der präzisen Erkennung der richtigen Bindungsstellen ab, was sich bei MSL2 in vivo, wo die Bindung ausschließlich an das X-Chromosom erfolgt, auch beobachten lässt. Zu diesem Zweck, und ähnlich wie bei anderen Transkriptionsfaktoren, muss MSL2 die wenigen funktionellen Bindungsstellen von einem großen Reservoir scheinbar ähnlicher, aber nicht-funktioneller Bindungsstellen unterscheiden. Bei in vitro Untersuchungen, bei denen MSL2 seine Bindungsstellen frei aus genomischer DNA auswählen kann, beträgt die Anreicherung der Bindungsstellen an das X-Chromosom jedoch nur etwa 50%, obwohl MSL2 immer noch das gleiche GA-reiche Bindungsmotiv bindet, das auch in vivo gefunden wurde.

Diese Arbeit zielt darauf ab, zu beschreiben, welche Faktoren fehlen, um eine getreue X-chromosomale Erkennung in vitro zu erreichen und um die Prozesse, die hinter der Selektion des Dosierungskompensationskomplexes im Besonderen und der Bindung von Transkriptionsfaktoren im Allgemeinen liegen, besser zu verstehen. Die Ergebnisse dieser Arbeit gliedern sich in drei Teile. Zuerst wird die „zellfreie Genomik“ beschrieben. Unter Verwendung eines Präblastodermextrakts von Drosophila-Embryonen (DREX) wird Chromatin auf genomischer Drosophila-DNA in vitro rekonstituiert, was dann als physiologisches Substrat für die Interaktion mit Transkriptionsfaktoren und Nukleosomen-Remodellierern verwendet werden kann. Zweitens werden MSL2 und ein bekanntes, mit MSL2 wechselwirkendes Protein, CLAMP (chromatin linked adapter for MSL proteins), in vitro auf Wechselwirkungen miteinander und mit dem DREX-assemblierten Chromatinsubstrat hin untersucht. Diese Proteine wurden in vitro bereits gut charakterisiert und bieten eine zuverlässige Kontrolle zum Testen von DREX-assembliertem Chromatin. Zellfreie Genomik ermöglicht die direkte Kontrolle der Konzentration relevanter Faktoren, was neue Einblicke in die Etablierung von X-chromosomaler Bindungsstellenselektion und in die Kooperation und Konkurrenz zwischen Transkriptionsfaktoren ermöglicht. Hier ist MSL2 als Modellprotein hervorragend geeignet, um zwischen produktiver und unproduktiver DNA-Bindung zu unterscheiden, da sich alle seine funktionellen Bindungsstellen auf dem X-Chromosom befinden. Schließlich wurden bioinformatische Analysen durchgeführt, die zeigen, wie komplexe DNA Formmerkmale beeinflussen wie Transkriptionsfaktoren ihre Ziele von einem Reservoir scheinbar ähnlicher, aber nicht-funktioneller Bindungsstellen unterscheiden. Für die MSL2-Bindungstellenselektion legen die Daten nahe, dass bestimmte „Shape"-Merkmale an funktionellen Stellen nicht als positives Erkennungssignal für MSL2 selbst dienen, sondern eher als negatives Selektionssignal gegen ein anderes häufig vorkommendes GA-bindendes Protein, GAF, das ansonsten MSL2 verdrängen würde. Diese negative Selektion demonstriert ein neuartiges Konzept mit breiterer Anwendbarkeit dafür, wie das Bindungsprofil eines Transkriptionsfaktors durch einen anderen, nicht verwandten Faktor mittels einer komplexen Sequenz- und „Shape"-Signatur geformt wird.

1. Introduction

1.1 Chromatin

All life is based on the information stored in strands of desoxyribonucleic acid [DNA, (Avery et al., 1944)]. The amount of stored information necessary to enable more complex life has led to increasingly long DNA strands, which needed to be stored within the cells. In all eukaryotes, this necessitates an enormous level of compaction. For example, a diploid human cell has to store about 6 billion base pairs of DNA on average, or 2 meters of DNA strand, in a nucleus of about 10 µm (Kornberg and Lorch, 1999). The biological structure required for this was first discovered by W. Flemming and was termed “chromatin” (from the Greek χρώμα, *chroma*, "colour") due to its ability to be strongly stained by certain dyes.

1.1.1 Nucleosome Structure

DNA compaction through chromatin is achieved in a multi-level process. At the first level of packaging, DNA is wrapped around histone octamers. The resulting nucleosome core particles consisting of DNA and histones are the fundamental repeating subunits of chromatin and resemble a thread wrapped twice around a spool (Kornberg, 1974). In total, the human genome is made up of about 30 million such nucleosomes. Nucleosomes are evolutionary very conserved and their function and structure is virtually identical for all eukaryotes. Structurally, all of the individual histone proteins have a very similar topology and consist of three alpha-helices which are arranged in a “histone fold” (Arents et al., 1991). Additionally, all histones have unstructured N- terminal domains termed “histone tails”, which can be modified. Histone H2A also has an additional C-terminal tail. Multiple modifications can be present at the tails including acetylation (Kuo and Allis, 1998), methylation, phosphorylation (Berger, 2002), ubiquitination (Zhang, 2003) and ADP-ribosylation (Ame et al., 2004). They are the basis for the gene regulatory functions of the chromatin and constitute a combinatory framework that regulates gene expression (Jenuwein and Allis, 2001).

The canonical octamer consists of two of each of the histone proteins H2A, H2B, H3 and H4 (Luger et al., 1997). To form the octamer, these proteins associate to form the heterodimers H2A- H2B and H3-H4 through so called “handshake structures”. Then H3-H4 dimers dimerize further to form an H3-H4 tetramer, which then further congregates into the octamer by binding of two H2A-H2B dimers on opposite ends of the tetramer. In absence of DNA H2A-H2B dimers and H3-H4 tetramers do not interact and form octamers only under high-salt conditions (Arents et al., 1991). This octamer is then complemented by 147 bp of DNA, which wraps around the core histone complex about 1.65 times (Andrews and Luger, 2011). The binding between histones and DNA is stabilized through interactions of the negatively charged phosphates of the DNA backbone and the positively charged surface of the histones. The line passing through the single central base pair of the nucleosome is called the dyad axis.

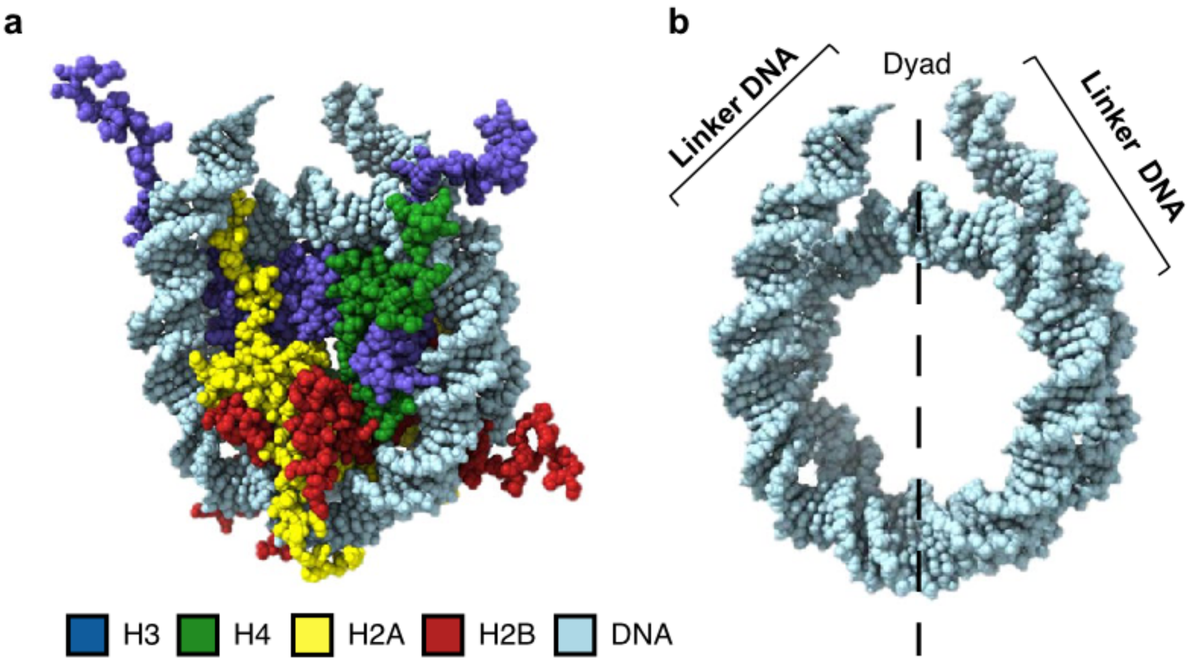


Figure 1: Nucleosome structure. Nucleosome disc view, model derived from PDB 1KX5 and PDB 1ZBB79. Colors represent the histone identity as indicated. b, Nucleosomal DNA and linker DNA (from PDB 1ZBB). Along the two-fold axis, nucleosomal DNA (145–147 bp) can be divided into two gyres (approximately 72 bp each). Linker DNA is the extra-nucleosomal DNA between two nucleosomes. The dyad is the center base of the nucleosomal DNA. Adapted and reprinted with permission from Nature (Zhou et al., 2019).

1.1.2 H1 in Chromatin Organization

The nucleosome only constitutes the first level of compaction of DNA within the nucleus. Its general architecture is shared by all eukaryotic organisms, while the folding of the chromosome is organized slightly differently in different organisms. For simplicity and because this work focusses on the Drosophila genome, mostly the proteins and mechanisms relevant in Drosophila will be discussed, unless indicated otherwise.

Along the whole genome fiber nucleosomes are assembled like “beads on a string” with higher levels of compaction upon association of the linker histone H1. The nucleosome bound by the H1 histone is termed a chromatosome. This association increases chromatin condensation by promoting formation of a 30-nm fiber (Bednar et al., 1998). Further compaction is achieved through association of non-histone proteins. Drosophila actually possesses two isoforms of H1. The BigH1 isoform is abundant during early embryogenesis and is later replaced by H1 during cellularization (Perez-Montero et al., 2013). This exchange was recently implicated in being essential for the maternal-to-zygotic transition (MZT) of gene expression (Schulz and Harrison, 2019). The nucleosome remodeler ISWI somehow promotes H1 integration into the chromatin as ISWI deletion leads to a strong reduction of H1 association to chromatin in vivo (Corona et al., 2007). In higher eukaryotes, the stoichiometry between nucleosomes and H1 is nearly 1:1 meaning that nearly all nucleosomes are bound by linker histones (Bates and Thomas, 1981). In vitro transcription assays using purified components for chromatin assembly revealed that H1 can reduce the basal transcription rate of RNA Polymerase II (Pol II) significantly. This reduction can, however, be overcome by transcriptional activators such as GAL4 (Laybourn and Kadonaga, 1991). Also, H1 addition increases nucleosome repeat length to 200-220 bp (Sandaltzopoulos et al., 1994). Additionally, H1 is necessary for the formation of all higher order chromatin structures such as pericentromeric heterochromatin and polytene chromosome structure (Lu et al., 2009).

1.1.3 Oligonucleosome Organization

Nucleosomal arrays are structurally well defined but the folding of the fiber is highly variable. Oligonucleosomes are organized in a zig zag or solenoid patterns and group into “clutches” interspaced by nucleosome-free regions (Krietenstein and Rando, 2020; Moraru and Schalch, 2019; Ricci et al., 2015). H1 was found to aid the formation of two-start helix conformations in chromatin (Garcia-Saez et al., 2018). Structural analysis suggested that a double helix-twisted tetranucleosome forms a fundamental unit of chromatin structure akin to the secondary structure found in proteins (Ohno et al., 2019; Song et al., 2014). At this level the oligonucleosomes in chromatin can assume a rigid so called 30-nm fiber, but nowadays it is considered to assume this conformation mostly in vitro (Eltsov et al., 2008; Tremethick, 2007). In the crowded environment of the nucleus, however, chromatin appears more like a “molten globule”, as neighboring fibers intercalate, disrupting the intrafiber contacts and forming the chromatin into a more uniform mass. Here polymer-polymer phase separation (PPPS) guides the formation of globular chromatin subcompartments with distinct molecular composition (Erdel and Rippe, 2018).

In mammals, chromatin is then further compacted through loop formation in which the cohesion complex binds chromatin and spools out loops until they encounter boundary factors such as CTCF (CCCTC-binding factor) (Fudenberg et al., 2017). This spatial arrangement leads to the formation of topologically associated domains (TADs) and further compaction, culminating in the chromosomal territories that encompasses the whole nucleus (Mirny, 2011). In Drosophila, CTCF is present, but not essential for TAD formation (Sexton et al., 2012). Instead, they require the transcription factor Zelda as a boundary factor between TADs. They are first demarcated after zygotic genome activation (ZGA) before which the genome is unstructured (Hug et al., 2017). High-resolution 3-dimensional mapping has revealed that the Drosophila genome harbors roughly 4000 such TADs (Wang et al., 2018).

1.1.4 Nucleosome Positioning and Sliding

Nucleosomes constitute physical and energetic barriers on the DNA for polymerases and transcription factors (Mavrich et al., 2008). Their exact position is therefore highly influential and highly regulated. Nucleosomes do not randomly position evenly throughout the genome but occupy sequences with the highest specific affinity for nucleosome binding. In vitro, they undergo confined diffusion in absence of remodelers to find the energetically optimal binding location (Rudnizky et al., 2019). The affinity of a histone octamers to a given sequence can vary over three orders of magnitude. Nucleosomes specifically prefer sequences with a 10 bp periodicity of TA dinucleotides (Thastrom et al., 1999). This conformation is favored because it faces the minor groove of highly bendable DNA (TA, AT) towards the histone at each of the DNA helical repeats (10 bp). On the other hand, poly(dA:dT) and poly(dG:dC) are particularly disfavored as they are intrinsically stiff and therefore inhibit nucleosome formation (McCall et al., 1985; Nelson et al., 1987). The nucleosome positioning maintained through the sequence alone is termed “rotational positioning” and can serve to predict nucleosome positions genome-wide with about 50% accuracy (Segal et al., 2006). These findings have been utilized to design nucleosome positioning sequences for in vitro experiments with well-defined nucleosome positions, such as the ‘601’ sequence (Lowary and Widom, 1998). In vivo, however, nucleosomes often do not position according to these rules. Using DNAse hypersensitivity assays, which show genomic locations of particularly low nucleosome occupancy in vivo, nucleosome-depleted regions (NDR) were found (Wu, 1980). These NDRs depend on the cell type and transcriptional state and are therefore not determined by sequence alone. This “translational positioning” is mediated through DNA binding factors or remodelers, which can slide nucleosomes to less favorable positions. Nucleosome depletion at transcription start (TSS) and termination sites (TTS) in vivo and at transcription factor biding sites is especially evident (Kaplan et al., 2009). Promotors are held nucleosome-free by multiple mechanisms, involving poly(dA:dT) sequences, specifically in yeast (Hughes et al., 2012), and this activity is independent of transcriptional activity. Additionally, the nucleosomes are arranged around TSS and TTS, forming NDRs at the promotors and precisely positioned +1 und -1 nucleosomes around these NDRs. The +1 nucleosomes are some of the most strongly positioned nucleosomes in the genome, though there do not seem to be any positioning sequences in this region (Zhang et al., 2009). Rather, the NDR is determined by the low affinity of histones to the DNA sequence and through binding of nucleosome remodelers such as the RSC complex (Remodeling the Structure of Chromatin) which then positions the +1 nucleosome (Wippo et al., 2011). The +1 nucleosomes then represent a barrier, against which further nucleosomes are aligned to by remodelers, such as ACF, resulting in a uniformly phased positioning pattern (Baldi et al., 2018b). The positioning precision of the nucleosomes gradually decreases downstream from there. Additionally, and specifically, at TSS nucleosomes undergo a constant cycle of assembly, remodeling and eviction through the activity of Pol II which increases the accessibility at these regions (Shivaswamy et al., 2008). In addition to the TSS, other DNA binding factors can act as such barriers to facilitate phased nucleosomal arrays such as general transcription factors (Oberbeckmann et al., 2021) and Su(Hw) or Phaser (Baldi et al., 2018b). Highly regularly positioned nucleosomes can facilitate the establishment and conservation of heterochromatin and gene silencing (Saxton and Rine, 2020). Nucleosomes are not only positioned around TSS but also slided throughout the entire genome. Nucleosome remodelers such as NURF, CHRAC or ACF (described in detail in 1.3.4) are responsible for this repositioning. Nucleosomes are moved in an ATP-dependend manner by weakening the DNA nucleosome interactions but without disrupting the core octamer (Hamiche et al., 1999; Längst et al., 1999). The DNA is displaced from the nucleosome at the edge, which leads to a bulge of DNA traveling around the nucleosome (Schiessel et al., 2001).

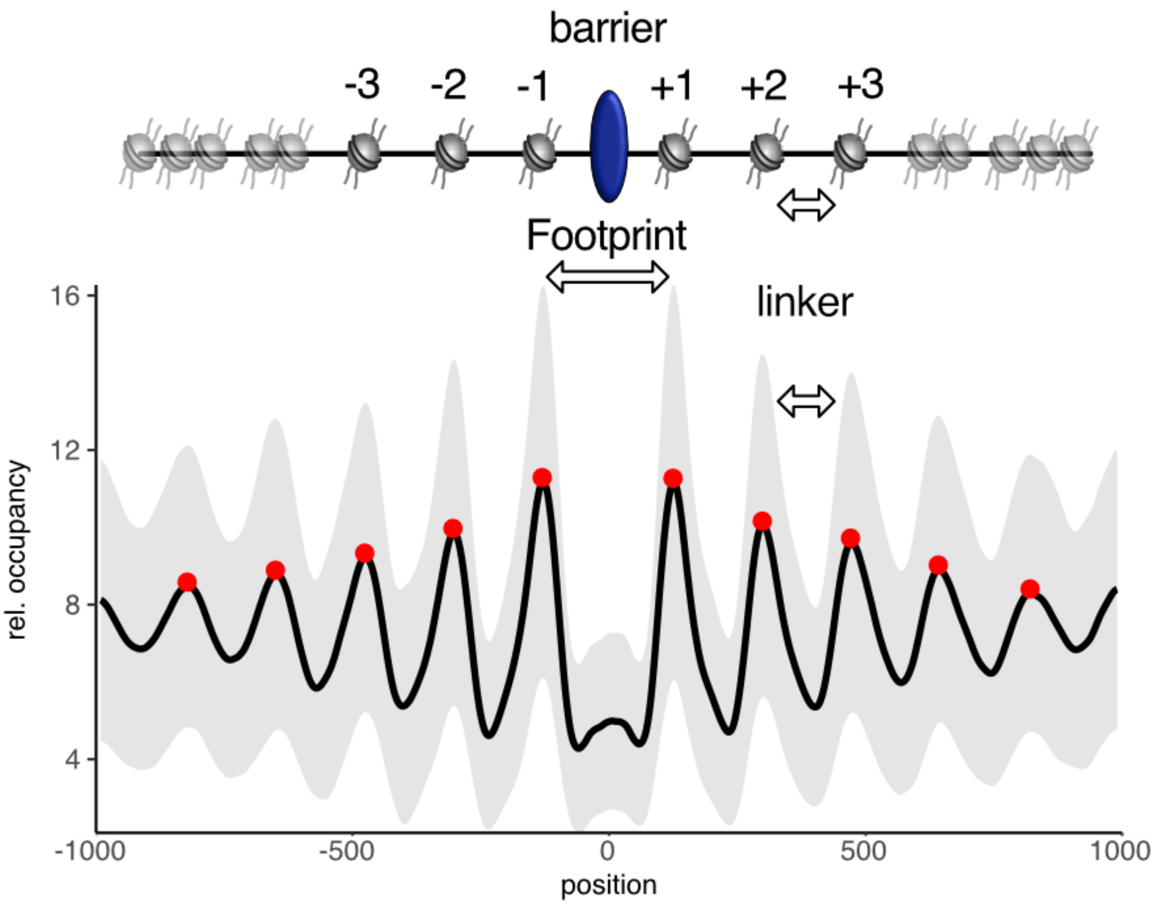


Figure 2: Phased nucleosomal array around a barrier site. Nucleosomal occupancy is determined by partial MNase digestion followed by sequencing (MNase-seq). Then occupancy windows around known barrier protein binding sites are cut computationally from the profile and cumulated. A transcription factor binds to the central motif, leaving a footprint in the nucleosome occupancy pattern. Then nucleosomes are shifted against it forming phased nucleosomal arrays. The red dots represent the nucleosome dyads.

This remodeling activity is often directed and recruited to certain locations by transcription factors (Kang et al., 2002; Xiao et al., 2001), but is also of general nature (Racki et al., 2009). The nucleosomes on chromatin are always in a dynamic state sliding around their preferred positions (Becker, 2002). This leads to a high permeability of chromatin for transcription factors at locations of strong sliding activity.

1.1.5 Histone Variants and Modifications

Nucleosomes exercise a high degree of regulatory control over the DNA organized in them generally through impeding DNA transcription (Brownell et al., 1996; Wasylyk and Chambon, 1979). This regulatory function is modulated through histone variants and histone tail modifications. If the “canonical” histones are replaced by sequence variants, the properties of the nucleosomes are modulated.

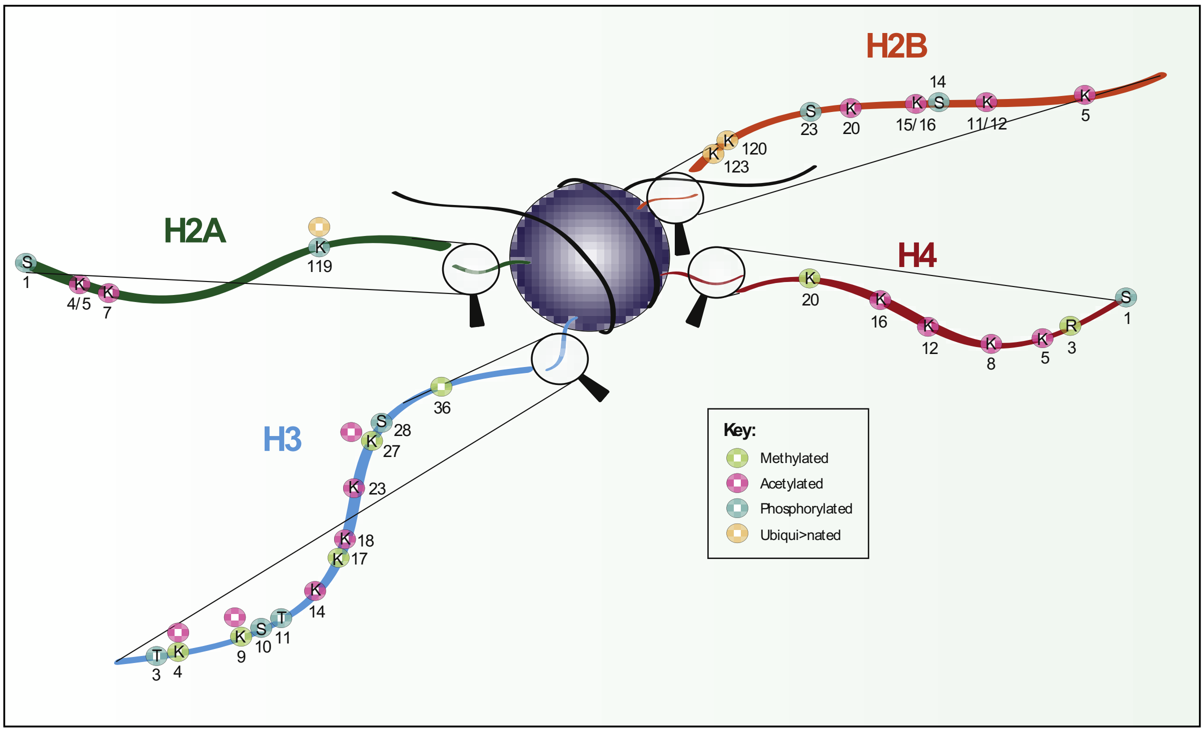


Figure 3: Post-Translational Modifications of the Histone Tails. The location of each modification is shown and the amino acid modified at each position is also annotated (K = lysine, R = arginine, S = serine, T = threonine). Colors depict how each residue is modified (green = methylated, pink = acetylated, turquoise = phosphorylated, beige = ubiquitinated). Adapted and reprinted with permission from Elsevier (Lawrence et al., 2016).

In Drosophila, the isoforms play important roles in a multitude of cellular processes like genome organization, transcriptional regulation and DNA repair (Henikoff and Smith, 2015; Talbert and Henikoff, 2010). They include, for example, nucleosomes containing CENPA (Centromere Protein A), an isoform of histone H3, which is exclusive to centromeres (Earnshaw and Rothfield, 1985). Additionally, alternative versions of histone H2A have been described, in Drosophila specifically this includes H2A.V, which combines the function of H2A.X and H2A.Z known in humans. In Drosophila it plays a role in promoter architecture and in DNA damage response (Baldi and Becker, 2013). The largest part of the regulatory function of histones is encoded on the histone tails, through chemical modifications. The histone variants and modifications present in the nucleosomes constitute a system, that regulates the transcriptional output of the underlying DNA. The modifications are added to histones post-transcriptionally by modifying enzymes and include acetylation, methylation, phosphorylation, ubiquitinylation, sumoylation and ADP-ribosylation. The combination of these modifications would, in theory, allow each histone in the genome to be uniquely modified (Zhao and Garcia, 2015). The modifications can affect transcription either by offering binding sites to interacting proteins or by modulating interactions between nucleosomes. H4K16 acetylation, for example, directly changes nucleosome interactions to reduce chromatin compaction and increase transcription (Shogren-Knaak et al., 2006). More often though chromatin modifications will indirectly affect transcription through the recruitment of effector proteins such as co-activators (or repressors) or remodeling complexes (Vettese-Dadey et al., 1996). For example, in flies histone 3 lysine 36 (H3K36) trimethylation is deposited by SET2 (Su(var)3-9, Enhancer-of-zeste and Trithorax-domain containing 2) and acts as a marker for active gene bodies. This modification then provides docking sites for several other proteins, such as nucleosome remodelers, like Isw1b in yeast (DiFiore et al., 2020) or chromodomain-containing proteins such as MSL3, discussed further in 1.4.2. Collectively, these modifications and the connected mechanisms ensure proper transcription by regulating the transcriptional output of the contained DNA.

Recent studies have shown that certain proteins can compartmentalize into droplets in a process termed “Liquid-Liquid-Phase-Separation” (LLPS). This concept helps to explain how biological functions can be partitioned into different sub compartments. This principle has also been demonstrated to be true for short nucleosomal arrays which undergo phase separation mediated though their tails (Gibson et al., 2019a). The significance of LLPS for multiple cellular processes including chromatin organization has also become obvious recently (Erdel and Rippe, 2018). The findings suggest that chromatin phase-separates into transcriptionally active or silenced chromatin compartments. LLPS, or in the case of longer constituting fragments such as chromatin PPPS, is thought to be promoted by the polymeric nature of the DNA fiber, the intrinsically disordered domains of chromatin proteins including histones and the macromolecular crowding within the cell nucleus (Narlikar, 2020).

1.1.6 Chromatin Types

Each chromosome is a single continuous DNA strand fully wrapped up in nucleosomes, however the density of the nucleosomes and the activity of the DNA differ throughout the genome. Historically, the chromatin along this filament was roughly divided into two types, the transcriptionally repressed heterochromatin and the active euchromatin (Olins and Olins, 2003). These areas are mostly defined through their accessibility for other DNA binding molecules, as nucleosomes occupy the DNA and therefore occlude DNA from other factors. In this binary system, euchromatin is the more lightly packed type and usually located more centrally in the nucleus. This area exhibits active transcription and is home to most of the constitutively active “housekeeping genes”. Heterochromatin is the more tightly packed type of chromatin and mostly located around the nuclear envelope. It is further subdivided into facultative and constitutive heterochromatin. Constitutive heterochromatin is the most compact form housing very few genes and is associated to the lamina. It is usually located around the centromere of each chromosome and constitutes large chunks of the Y-chromosomal DNA in most species. Facultative heterochromatin, on the other hand, plays a role in gene regulation and can transition between hetero- and euchromatic states. The genes housed here are silenced or expressed depending on the state of the surrounding chromatin. Facultative heterochromatin formation is mostly regulated though polycomb group proteins, which can bind and spread over chromatin depending on the presence of certain histone variants or modifications (Bantignies et al., 2011).

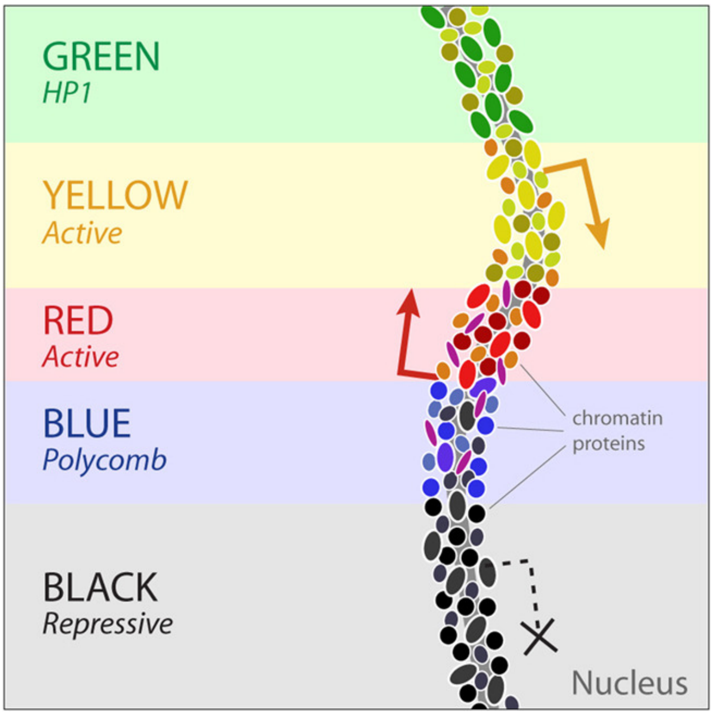


Figure 4: Different chromatin regions can be grouped into 5 different types. These types are characterized by the pattern of histone modifications and the presence of structural proteins such as HP1 and Polycomb. Reprinted with permission from Elsevier (Filion et al., 2010).

Polycomb-repressed domains engage in long-range contacts with each other, influencing genome organization leading to the formation of repressed compartments (Sexton et al., 2012). More recently, by looking at the combinatorial pattern of certain histone modifications and chromatin-associated proteins, chromatin has been classified into 5 (Filion et al., 2010) or 9 (Kharchenko et al., 2011) chromatin types or “colors”.

In the five-color system YELLOW and RED represent the euchromatin fraction. The chromatin around transcription start sites (TSS) of ubiquitously expressed housekeeping genes are marked YELLOW while the TSS of more restrictively expressed genes that are linked to specific tissues or developmental processes are marked in RED. BLUE and GREEN represent the heterochromatic fraction. GREEN is characterized by HP-1 and related proteins and is considered to be constitutive heterochromatin. BLUE, on the other hand, is characterized by the presence of polycomb group (PcG) proteins and is considered to be facultative heterochromatin. BLUE chromatin is mostly developmentally repressed. BLACK chromatin is mostly silenced and featureless, but can be activated in certain tissues, indicating tightly restricted, but nonetheless dynamically regulated regions (Filion et al., 2010).

Further combinatorial analysis of the histone modifications has then led to a 9-chromatin states model: State 1 is marked by H3K36me3/me2 and H3K9ac and represents active promotors. State 2 is marked by H3K36me3, important for transcription elongation, and represents the exonic regions. State 3 is marked by H3K27ac, H3K4me1 and H3K18ac and represents enhancer regions. State 4 is marked by H3K26me1 but notably lacking H3K27ac and is also present in intronic regions. State 5 is marked by H4K16ac and H3K36me3 and is notably enriched on the X chromosome in male cells. This state is associated with the dosage compensation, which is explained in detail in chapter 1.4. State 6 is marked by Polycomb-mediated repression. State 7 and 8 are marked by high or medium levels of H3K9me2/me3 and represent different types of heterochromatin. Lastly, State 9 represents extensive silent domains and serves as a “background” state (Kharchenko et al., 2011).

1.1.7 DNA Shape

The DNA molecule is a long polymer of nucleotides which forms a double-strand helix (Watson and Crick, 1953). The backbone of this helix consists of the alternating ribose and phosphate groups, while the central part is made up of the planarly stacked nucleotide bases. The two helical chains are interconnected by hydrogen bonds formed between a purine and a pyrimidine, usually in the specific pairs of cytosine and guanosine or adenine and thymine, respectively. The strands are further stabilized by the base pair stacking interactions between the aromatic nucleobases (Danilov and Tolokh, 1984). The DNA double helix is flexible and can adapt a variety of conformations through supercoiling and bending, which influences the affinity of interacting proteins (Irobalieva et al., 2015). Especially the conformation, in which the nucleotide bases are stacked, can vary extensively depending on the sequence and influence DNA-protein interactions. Moreover, the DNA conformations or “shapes” do not only depend on the specific bases making up the strand at each position, but also on the neighboring bases around each position. The main structural properties influencing DNA-protein interactions are Minor Groove Width (MGW), Propeller Twist (ProT) and DNA Roll (Roll) (Gordan et al., 2013; Rohs et al., 2009) (Figure 5). The MGW describes the asymmetry of the helical strands opposing each other. The grooves formed between the DNA backbones are unequal. The wider or major groove is 2.2 nm wide and while the smaller or minor groove is just 1.2 nm wide. Within the major groove the base pair sequence is more readily accessible to most sequence-specific DNA binding factors. ProT and Roll describe the conformations of the base pair stacking in regard to each other. The Propeller Twist describes the rotation of the partnering bases within a base pair to each other, while the DNA Roll describes the degree of rotation between neighboring base pairs (Figure 5). These structural properties influence TF binding and including them into binding site predictions outperforms PWMs. The complementary nature of DNA shape and sequence in determining TF binding has been demonstrated in vitro by numerous groups (Joshi et al., 2007; Rohs et al., 2009).

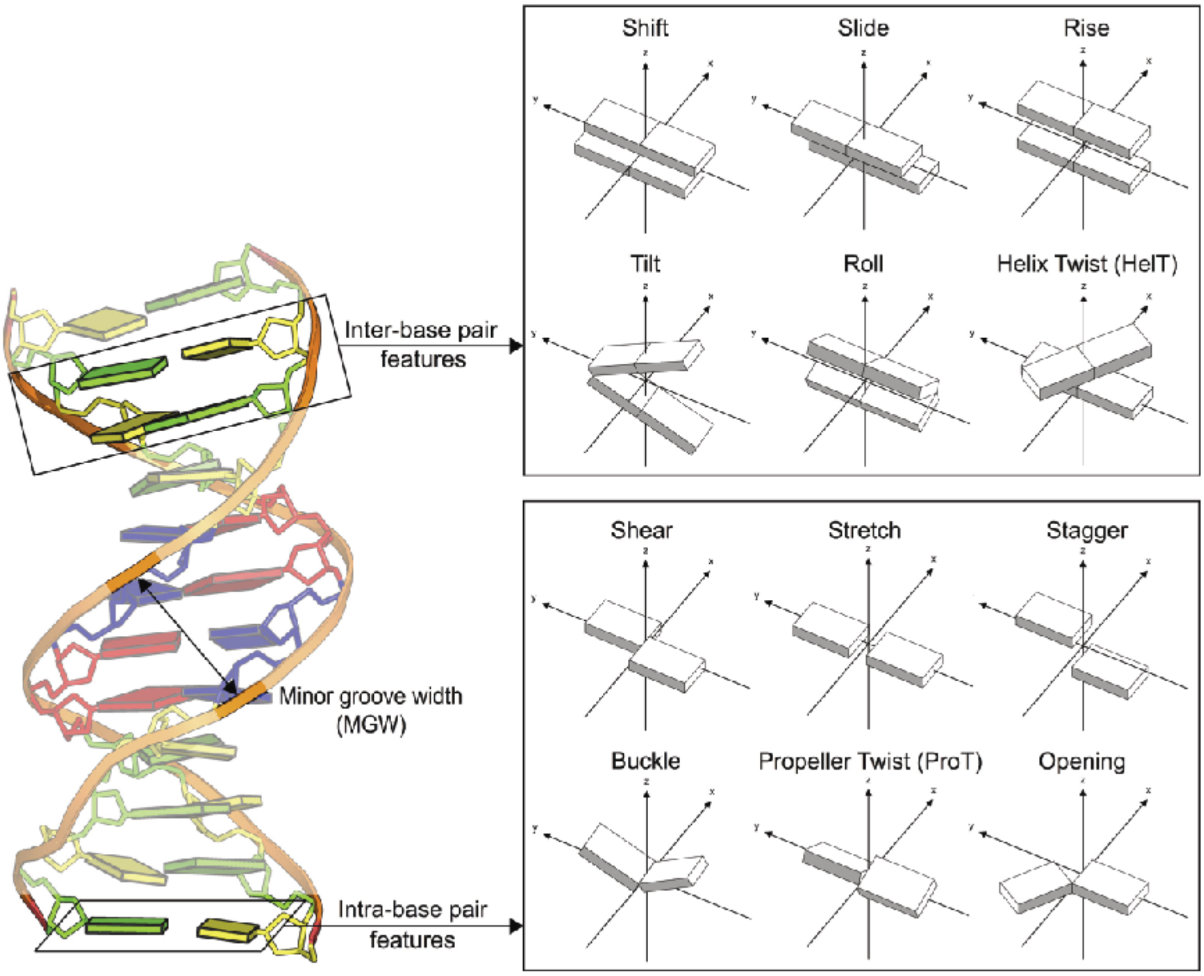


Figure 5: Representation of 13 DNA shape features. Schematic representation of a DNA fragment (PDB ID: 1BNA taken from the Protein Data Bank) with definition of MGW, inter-base pair and intra-base pair parameters. Adapted and reprinted with permission from Oxford University Press (Li et al., 2017).

1.2 Reconstitution of Chromatin in vitro

1.2.1 Minimal Reconstitution Systems

To study molecular mechanisms of protein-DNA interactions it is preferable to use chromatin over DNA as DNA-binding factors have evolved along with chromatin. Reconstitution in vitro can be achieved by basically two different approaches, an ATP-independent deposition of nucleosomes onto DNA and ATP-dependent assembly of nucleosomal arrays (Lusser and Kadonaga, 2004).

The simplest and ATP-independent approach to assemble nucleosomal arrays is through salt gradient dialysis (SGD). In this approach histone octamers are stabilized in 2 M NaCl. DNA is added to the sample and the salt concentration is slowly reduced to physiological levels, depositing nucleosomes onto the DNA in the process. During this, first H3/H4 tetramers are deposited at about 1 M NaCl, followed by the addition of H2A/H2B at about 600 mM. If the pure core histones used were expressed as recombinant proteins this has the advantage that they are free of any posttranslational modifications and devoid of contaminating factors. Additionally, this allows for easy incorporation of mutant or chemically modified forms. Unfortunately, the nucleosomes in this approach will not adopt physiological positions but positioning will be guided by energy minimization of the system. Negatively supercoiled, circular DNA is best suited for this kind of reconstitution, as every added nucleosome absorbs a negative supercoil from to the substrate by converting a plectonemic, negative supercoil into a toroidal supercoil, wrapping the DNA around the octamers. Reconstitution is more efficient in this case as compared to linear DNA (Pfaffle and Jackson, 1990). As the nucleosomes are not deposited randomly but according to local energetic minima (Kaplan et al., 2009; Segal et al., 2006), utilizing the DNA sequence-specific affinities can be used to place the nucleosomes on well-defined positions. The most common sequence used for this DNA directed positioning is called “601” (Lowary and Widom, 1998).

It is also possible to assemble chromatin in vitro in absence of assembly factors by mediation through high negative-charge density molecules such as of Poly(L-glutamic-acid) (Stein et al., 1979) or acidic polysaccharide pectin (Sobolewski et al., 1993). These molecules stabilize the core histones by forming a large complex with them and can in turn be replaced by DNA (Oohara et al., 1983). This process deposits the octamers onto the DNA fiber at energetic minima in a similar way as salt gradient dialysis.

Another possibility to assemble chromatin is through the use of ATP-dependent assembly systems. The most minimalistic version of this uses purified factors such as ATP-dependent chromatin assembly factor (ACF) and the histone chaperone NAP-1 to ease assembly (Fyodorov and Kadonaga, 2003), while more complex systems utilize crude cell extracts from Drosophila embryos (Becker et al., 1994) or Xenopus oocytes (Glikin et al., 1984).

In the minimalistic approach histones are deposited onto DNA in vitro by an ATP-dependent assembly machinery in which ACF and NAP-1 function synergistically in the assembly of nucleosomes (Ito et al., 1997). At first, histones are bound by the histone chaperone NAP1. NAP1 binds directly to the free histones and aides assembly by blocking non-nucleosomal histone-DNA interactions (Andrews et al., 2010). Then, nucleosomes are assembled onto a plasmid in an ATP-dependent manner utilizing the assembly ability of a general chromatin remodeler ACF (Ito et al., 1997). ACF itself consists of an ISWI ATPase and an Acf1 subunit, both of which are required for its function (Ito et al., 1999). In the presence of ATP, ACF aids assembly of nucleosomes onto the DNA and modulates the nucleosome spacing. Specifically, ACF increases the distances between nucleosomes and spaces them more regularly. The formation of these regular nucleosomal arrays inhibits transcription (Fyodorov et al., 2004). Experiments testing the expression from nucleosomal arrays on plasmids using this system indicate that the resulting nucleosome positioning and mobility are physiological (Ito et al., 1997). After assembly, H1 can be incorporated, yielding chromatosome arrays with physiological properties (Fyodorov and Kadonaga, 2003).

1.2.2 DREX-Reconstituted Chromatin

A different approach to assemble chromatin utilizes whole-cell extracts, which contain all relevant factors necessary for assembly. During their earliest development, Drosophila embryos undergo several rounds of rapid DNA replication while forming a polynucleated syncitcum. To facilitate this, the embryos already contain large stocks of maternally-deposited histones and chromatin assembly factors (Becker and Wu, 1992). In this extract, the excess histones are sequestered from the cytoplasm by anchoring them to lipid droplets for storage (Li et al., 2012). This serves as a buffer for H2A.V and H2A concentrations and prevents turnover of excess histone protein (Li et al., 2014; Stephenson et al., 2021). Drosophila embryo extracts were first used in 1979 to assemble chromatin like structures in vitro (Nelson et al., 1979) and have since been extensively used in studies of DNA replication (Crevel and Cotterill, 1991), in vitro transcription (Kamakaka and Kadonaga, 1994; Kamakaka et al., 1991) and most notably in vitro chromatin assembly (Becker et al., 1994; Becker and Wu, 1992). Chromatin reconstituted using the extracts from preblastoderm Drosophila embryos (roughly 90 minutes after egg laying) efficiently assemble DNA into complex chromatin (Becker and Wu, 1992). This chromatin harbors a large physiological proteome containing hundreds of proteins (Voelker-Albert et al., 2016), including nucleosome remodeling factors such as ACF (Ito et al., 1997), CHRAC (Alexiadis et al., 1998; Varga-Weisz et al., 1997) or NURF (Tsukiyama and Wu, 1995) and bound insulator complexes such as Su(Hw) and Phaser (Baldi et al., 2018a). This leads to the formation of chromatin that exhibits physiological nucleosome spacing and shows phased nucleosomal arrays adjacent to tightly bound "barriers" proteins such as Su(Hw) or Phaser. The formed chromatin is highly dynamic and transcription factors are able to access even nucleosome-occupied target sites through the activity of the present nucleosome remodelers (Wall et al., 1995). This allows for physiological transcriptional activation of chromatinized genes in this in vitro system (Kamakaka et al., 1991; Sandaltzopoulos et al., 1994). DREX-assembled chromatin also incorporates H2A.V into the nucleosomes. This is noteworthy, as this in vitro system can phosphorylate this histone variant and utilize it as a damage sensor in a in vivo-like manner. This demonstrates that DREX chromatin is able to utilize complex signaling pathways in a cell-free environment (Harpprecht et al., 2019). Chromatin assembled this way is notably free of H1 as the linker histone is only expressed after the beginning of ZGA. Instead, DREX-assembled chromatin resembles the early preblastoderm state of chromatin characterized by the presence of BigH1 (Henn et al., 2020; Perez-Montero et al., 2013) and HMG-D (Ner and Travers, 1994). These variants usually regulate the zygotic genome activation in vivo. Nonetheless, H1 can be incorporated into DREX-assembled chromatin if added during the assembly and readily replaces the BigH1 and HMG-D leading to a change in linker length and physiological degrees of transcriptional repression (Becker and Wu, 1992). Chromatin assembly extracts can be applied to chromatinize any source of DNA including mega-base-pair long linear fragments (Becker et al., 1994; Climent-Canto et al., 2020), which is difficult to achieve with ATP-independent approaches. Complexity is, however, also a disadvantage, as it can complicate the interpretation of experimental findings.

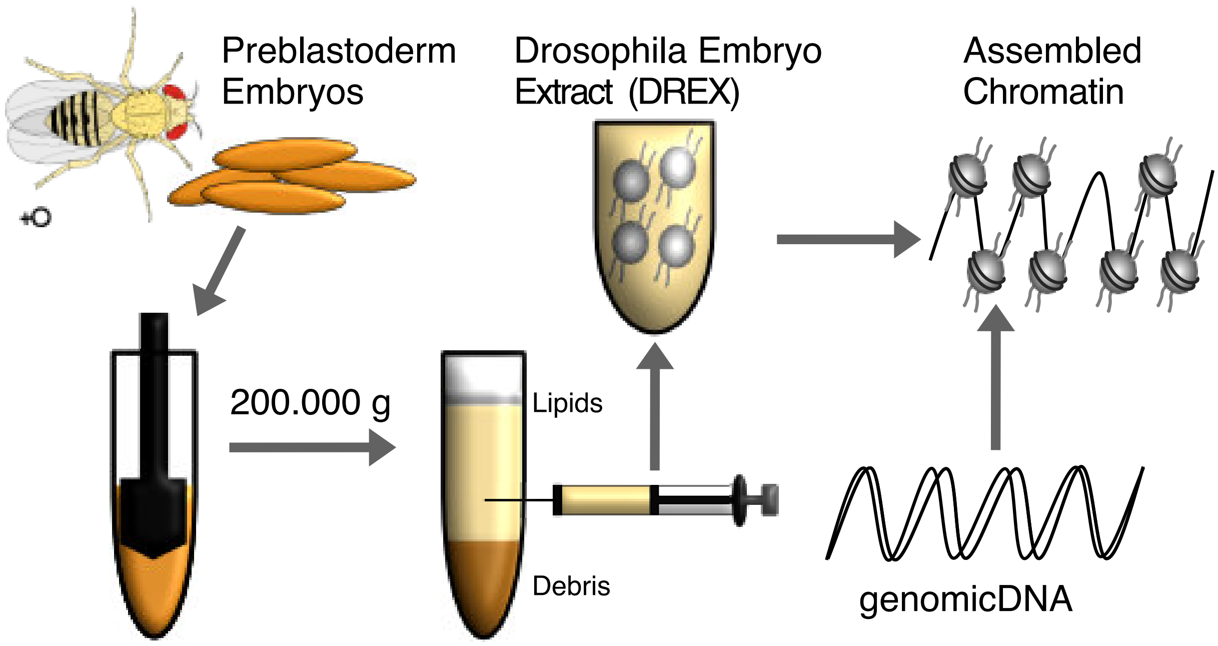


Figure 6: Schematic depiction of the chromatin reconstitution workflow using DREX extracts. Preblastoderm embryos are collected and disrupted using a pestle. Then cell debris and lipids are separated from the DREX using centrifugation. The resulting extract can be added to any source of DNA and will assemble physiological chromatin.

It was recently shown that DREX-assembled chromatin condenses into aggregates sized roughly from 1-10 µm (Eggers and Becker, 2021). As this work utilizes ~150 kbp long fragments , it can be assumed that this condensation is governed by the same polymer-polymer phase separation dynamics as chromatin is in vivo (Erdel and Rippe, 2018; Gibson et al., 2019b; Maeshima et al., 2016). Being folded by the same physical rules, one can assume that the local chromatin concentrations within these condensates approach the physiological values present in the nuclei. Concentrations of chromatin obtained through other approaches is often concentrated orders of magnitude lower. The addition of recombinant TFs to previously assembled genomes mimics to some extent the process of the 'zygotic genome activation' (ZGA), when the first wave of transcription leads to functional diversification of the naïve preblastoderm chromatin (Hamm and Harrison, 2018). DREX-assembled chromatin has been used before to study TF interactions in vitro, but not in genome-wide reconstitutes (Wall et al., 1995). The approach used by me expands these earlier findings.

1.3 Chromatin Interacting Factors

1.3.1 Transcription Factors

Transcription factors (TFs) are proteins that bind DNA and influence the rate by which genes are transcribed. They are necessary to regulate the transcription rates in different tissues at different times through development and according to external stimuli (Arendt et al., 2016;

Lambert et al., 2018; Lee and Young, 2013; Spitz and Furlong, 2012). Their function is usually targeted to promoters or enhancers by sequence-specific binding to short DNA sequence motifs that display a certain consensus sequence (Yan et al., 2021). Variations of that sequence can modulate the affinity of the TF, as can the nucleosome organization (Mirny, 2010). The sequence motif with the highest affinity for TFs can be determined by DNA or chromatin immunoprecipitation and sequencing of the associated DNA (DIP-seq / ChIP-seq) (Gossett and Lieb, 2008; Jolma and Taipale, 2011; Liu et al., 2005). They are most commonly visualized in position weight matrices (PWMs), which show the relative contribution of each base at each position of the sequence and scales it to the information content present at that position. Common domains in transcription factors that mediate the DNA binding are leucin zippers or homeobox domains, and most often so-called zinc-finger domains. After translation, any TF needs to be imported into the nucleus and must scan the entire genome and then bind only a small subset of physiologically relevant loci while ignoring thousands of other highly similar but essentially nonfunctional sites, which we term “decoy” sites . A typical metazoan genome contains thousands of potential binding sequences that match the TFs PWM, but only a small fraction of these is actually bound. However, many transcription factors are promiscuous binders and localize also to regions with low affinity and no regulatory function (Paris et al., 2013). One explanation of how potential but nonfunctional sites are excluded by TFs is that most of the potential binding sites are wrapped up in nucleosomes. The nucleosome represents a considerable barrier to any transcription factor-DNA interaction (Makowski et al., 2020). Nucleosomes hinder TF binding sterically and occupancy through nucleosomes is indeed a strong competitor of TFs for DNA binding (Workman et al., 1988). Transcription factors must compete with nucleosome assembly or cooperate with nucleosome remodeling factors to integrate themselves into the chromatin landscape (Morgunova and Taipale, 2017).

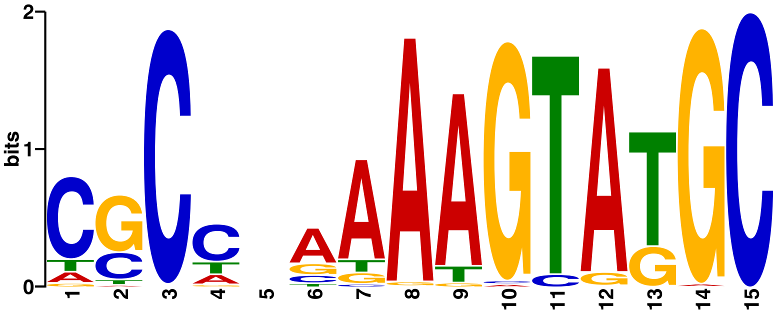


Figure 7: Typical position weight matrix (PWM) of a transcription factor binding site. The numbering underneath shows the position within the binding motif, the letters indicate the relative probability of base occurrence at each position, while the total height of all letters in a column combined indicates the information content of each position. In this case the Su(Hw) motif was derived from 2257 known binding sites and determined by MEME.

Only so-called “pioneer factors” are able to bind their targets regardless of nucleosome position. All others have to compete for binding with the nucleosomes and usually depend on nucleosome remodeling factors to actively free their binding sites (Teif and Rippe, 2009). The DNA can, however, become accessible to TFs as the nucleosomes can spontaneously unwrap from the DNA fiber (Li et al., 2005b) or chromatin remodelers or pioneering factors may specifically enhance chromatin accessibility around binding sites (Miller and Widom, 2003). Increasingly also the DNA shape is appreciated as a subtle discriminator of TF-DNA interactions (1.1.7). Certain shape features flanking a target sequence can reduce the speed of transcription factor diffusion through the genome (Mathelier et al., 2016; Suter, 2020; Yang et al., 2017). The intrinsic specificity of the DNA binding domains may be allosterically modulated by assembly into protein complexes or by cooperative interactions with other TFs.

1.3.2 Pioneering Factors

If a transcription factor can bind directly and independently of other factors to condensed chromatin, it is considered a pioneer (Zaret, 2020; Zaret and Carroll, 2011). These factors were first discovered using in vivo footprinting to determine which factors bind first to specific enhancers (McPherson et al., 1993). Pioneer factors can bind to silenced chromatin, which is inaccessible to other non-pioneering transcription factors (Zaret et al., 2016). Their binding often precedes the binding of other factors, which get recruited to the site through the pioneer's action. This allows factors like nucleosome remodelers or TFs to bind chromatin, which in turn can have repressive or supportive effects on transcription, but generally they increase local chromatin accessibility (Cirillo et al., 2002). Cooperativity between different factors just to outcompete nucleosome binding is not considered pioneering (Zaret and Carroll, 2011). Notably, pioneering factors are not dependent on nucleosome remodeler action or ATP to bind their targets. They bind to the outside face of nucleosomal DNA from the edge to the dyad axis center (Zhu et al., 2018) and some pioneers are even able to displace the linker histone to open up binding sites for other factors (Kagawa and Kurumizaka, 2021; Taube et al., 2010). While scanning the chromatin fiber for their target sequence their mobility is drastically lower than that of other transcription factors, as affinity for nucleosomal DNA of pioneer factors is much higher (Nakahashi et al., 2013; Sekiya et al., 2009).

Pioneering factors may affect nucleosome structure either directly or indirectly. They may directly perturb nucleosome-DNA interactions in such a way that they unwrap parts of the DNA from the nucleosome (Donovan et al., 2019) or evict the nucleosome from the fiber entirely (Laptenko et al., 2011), thus opening up potential binding sites for other factors. Pioneering factors can also disturb inter-nucleosomal interaction by repositioning of the histone tails, which are necessary for higher-order nucleosome stacking (Dodonova et al., 2020). They can also act indirectly by recruiting ATP-dependent chromatin remodeling factors (Yan et al., 2018) leading to remodeler-dependent nucleosomal arrays around their binding sites.

1.3.3 Factor Cooperativity

To be able to bind their target DNA sequences most transcriptions factors need binding site to be available and unoccupied by histones (Almouzni et al., 1990; Almouzni and Wolffe, 1993; Ramachandran and Henikoff, 2016). In vivo, transcription factors could theoretically bind the DNA after replication before histones occupy the DNA, but it was shown that histones outcompete TFs after the passage of the replication fork and TFs then have to compete for binding sites against histones thereafter (Ramachandran and Henikoff, 2016; Vasseur et al., 2016).

To compete, many TFs work in concert either by direct interaction and formation of homo- and heterotypic multimeric complexes (Espinas et al., 1999) or by indirect cooperativity. In direct interaction the proteins physically interact and stabilize each other's binding. For indirect cooperativity to take effect, competition against nucleosomes is necessary. This cooperativity is especially strong if the two binding sites are within the same nucleosome and do not sterically hinder each other (Moyle-Heyrman et al., 2011). In eukaryotes, most transcriptional regulation is conferred not by a single factor but by combinations of multiple transcription factors binding in close proximity to cis-regulatory elements or by the factors interacting with multiprotein complexes or chromatin remodelers. Only this complexity allows to code enhancers and promotors in a way that leads to appropriate expression for each cell type and developmental stage. TFs binding close to each other enhance each other's respective binding even in absence of direct interaction, as they collectively compete against histones. Cooperativity is a widespread mechanism to outcompete histone-DNA interactions (Hebbar and Archer, 2007; Sönmezer et al., 2020). It is especially strong if the distance between binding sites is "in phase" with the DNA fiber rotation (Liu et al., 2020). Also, the DNA sequences in the vicinity of binding site can influence TF affinities, through recognition of certain sequence symmetries (Afek et al., 2014).

In addition to this TF-TF cooperativity, transcription factors may recruit ATP-dependent remodeling complexes (Cosma et al., 1999) and cofactors, such as histone acetyl transferases (HAT), to overcome the barriers presented by chromatin (Narlikar et al., 2002).

1.3.4 Nucleosome Remodeling Factors

Nucleosome remodelers are protein complexes that utilize ATP to slide or evict nucleosomes from the underlying DNA fiber in order to regulate gene expression and DNA accessibility (Becker and Horz, 2002). Based on their structure and function the remodelers can be grouped into four groups, the SWI/SNF (Switch/Sucrose-Non-Fermenting), ISWI (Imitation SWI), CHD (Chromodomain Helicase DNA-binding) and INO80 (Inositol Requiring 80) protein families (Hargreaves and Crabtree, 2011; Tyagi et al., 2016). All families have an ATPase subunit from the SNF2 helicases (superfamily 2 DEAD/H-box)(Bork and Koonin, 1993). They differ, however, in the number of subunits from 2 in certain ISWI complexes to 11 in SWI/SNF complexes (Kingston and Narlikar, 1999).

Remodelers have been implicated extensively in the regulation of transcription initiation, elongation and termination as well as in histone variant deposition and DNA repair (Tyagi et al., 2016). Remodeling is constantly active and inhibition of remodelers reduces TF binding within minutes (Iurlaro et al., 2021). Molecularly, they all function similarly by hydrolyzing ATP to introduce superhelical torsion into DNA, which then affects the DNA-histone interactions in nucleosomes (Havas et al., 2000). This usually leads to a change in position of the nucleosome, which can either increase or decrease the accessibility of a given site for other potential DNA binding proteins. Therefore, nucleosome remodeling can facilitate both activation or repression of genes (Becker and Workman, 2013; Korber and Becker, 2010; Tyler and Kadonaga, 1999).

Noteworthy for this thesis are specifically the ISWI-type nucleosome remodelers contained abundantly in DREX: ACF, CHRAC and NURF (Ito et al., 1997; Längst et al., 1999; Tsukiyama and Wu, 1995; Varga-Weisz et al., 1997).

The related CHRAC and ACF remodelers act on a general level, roaming the whole genome and moving nucleosomes in a way to evenly space them out (Racki et al., 2009), but it has also been seen to localize to heterochromatic regions (Machida et al., 2018). The regular nucleosomal spacing by ACF is thought to facilitate the formation of repressive chromatin structures (Fyodorov et al., 2004). Others, such as the SWI/SNF complex, act in a more targeted manner (Judd et al., 2021). Such targeted action is typically seen by transcription factors and indeed TFs interact with and recruit remodeling complexes to their binding sites (Kang et al., 2002). But as many transcription factors also depend on remodelers to access their binding sites (Iurlaro et al., 2021), a scenario wherein both factors in turn recruit each other is probable.

NURF is a remodeling factor that slides nucleosomes and interacts with multiple transcription factors to enhance accessibility around their respective binding sites, such as the GAGA Factor (GAF)(Judd et al., 2021; Kang et al., 2002; Xiao et al., 2001). Importantly for this thesis, NURF is also recruited to sites bound by CLAMP, a known interactor of the Dosage Compensation Complex (Urban et al., 2017a). This cooperation enhances the accessibility of the male X-chromosome for the DCC. Loss of NURF has been shown to be detrimental to the compensation process and disrupts the global organization of the whole chromosome (Lucchesi et al., 2005).

1.4 Dosage Compensation in Drosophila

1.4.1 Dosage Compensation

Dosage Compensation is a process in many species in which the transcriptional output between the biological sexes is equalized. This is necessary because sex is often determined by heteromorphic sex chromosomes. As sex chromosomes harbor not only genes necessary for sex-determination but many other genes, the imbalance of expression can be deleterious for the heterogametic sex (having two different sex chromosomes)(Lucchesi, 1978). For example, in flies and in humans, sex is determined by the presence of the Y chromosome in males next to a single X chromosome, while females have two X chromosomes. In birds, on the other hand, sex is determined by a Z and W chromosome in females and two similar Z chromosomes in males (Lucchesi, 2018). Different systems for equalization have developed through evolution and are grouped as dosage compensation, but they all have in common that they use chromatin-based mechanisms to regulate transcription (Lucchesi, 1978; Lucchesi et al., 2005). In humans, for example, females harbour two X chromosomes as opposed to a single one in males. Here, one of the X chromosomes is inactivated through the long, noncoding RNA Xist pathway, leading to a single transcriptionally active X chromosome in males and females (Chow et al., 2005).

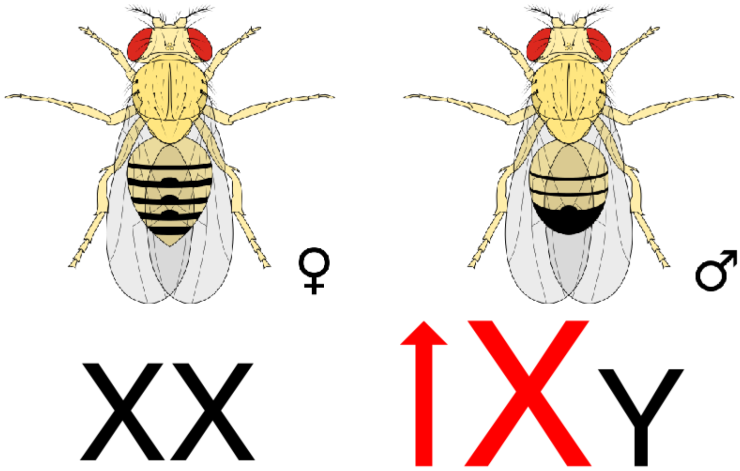


Figure 8: Dosage compensation in flies. The single X-chromosome in male flies is upregulated about 2-fold to equalize the transcriptional output of X chromosomal genes between the sexes.

In Drosophila, the same types of sex chromosomes are present. But in contrast to humans, Drosophila has developed a compensation system, in which the transcriptional output of most genes on the single X chromosome in males is approximately doubled, leading to equal expression throughout the sexes. (Straub and Becker, 2007).

1.4.2 Components of Drosophila Dosage Compensation Complex

In Drosophila males, X chromosome dosage compensation is achieved by the ribonucleoprotein Male-Specific-Lethal Dosage Compensation Complex (MSL-DCC). This complex consists of 5 proteins: MSL1 (Male-Specific-Lethal 1), MSL2, MSL3, MLE (Maleless) and MOF (Males-absent-on-the-first). The complex also incorporates one of two redundant long, noncoding RNAs named roX1 (RNA-on-the-X 1) and roX2. If dosage compensation is in any way disrupted in males this leads to male-specific lethality, while the artificial expression of MSL-2 in females, which activates this pathway, leads to low viability, sterility and developmental delays (Belote and Lucchesi, 1980; Kelley et al., 1995).

MSL1 is a scaffolding protein harboring interaction sites for the other subunits of the complex, except MLE. While the central coiled-coil domain contains interaction domains for MSL2 (Li et al., 2005a), the C-terminal PEHE domain interacts with MSL3 and MOF (Morales et al., 2004). Also, MSL1 is thought to mediate dimerization of the whole complex through homodimerization of MSL1 through its coiled coil domain (Hallacli et al., 2012). MSL3 contains a chromodomain necessary to bind H3K36me3, which is a common marker for actively transcribed chromatin and is thought to target DCC to substrate chromatin (Larschan et al., 2007).

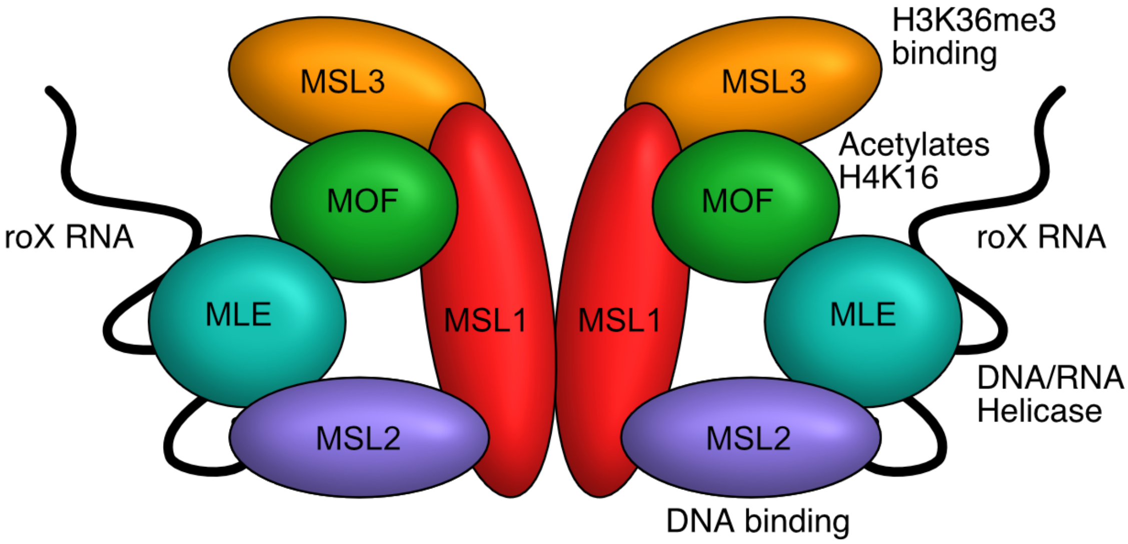


Figure 9: Graphical representation of DCC structure. The complex consists of 5 proteins (3 MSL proteins, MOF, MLE) and a long, noncoding RNA (roX). MSL2 is responsible for DNA binding and acts as an E3 ubiquitin ligase, MLE is an ATP-dependent RNA helicase, MOF acetylates H4K16 and MSL3 binds H3K36me3 through its chromodomain.

MLE is an ATP-dependent DEAD-box RNA helicase able to unwind double-stranded RNA (Lee et al., 1997) and is known to associate with the roX RNA (Meller et al., 2000). The roX RNAs differ greatly in size, sequence composition and expression patterns during development (Amrein and Axel, 1997; Meller, 2003; Meller et al., 1997) yet seem to be mostly redundant. RoX1 RNA is expressed earlier in development and in both sexes, while roX2 is expressed later and only in males (Meller, 2003). MOF is an acetyltransferase, the effective writer module of the DCC, and can acetylate H4K16 through its C-terminal HAT domain (Smith et al., 2000). This leads to transcriptional activation of targeted genes (Akhtar and Becker, 2000).

MSL2 is the only male-specific component of the complex and mediates the targeting and DNA binding. Its translation is repressed in females through the sex-lethal translational regulator SXL (Bashaw and Baker, 1997). MSL2 is involved in activation of roX2 transcription, without which the complex cannot be assembled (Bai et al., 2004; Rattner and Meller, 2004). It contains the C-terminal CXC and a proline/basic-residue rich (Pro-Bas) domain necessary for specific and general DNA binding, respectively, and an N-terminal RING domain. This domain works as an E3 ubiquitin ligase, shown to auto-ubiquitinylate MSL2 and the other components of the complex to maintain proper stoichiometry (Schunter et al., 2017; Villa et al., 2012). This domain also mediates MSL1 interaction (Copps et al., 1998). Additionally, MSL2 harbors a domain for interaction with CLAMP (chromatin linked adaptor of MSL proteins), the CLAMP binding domain (CBD). Most recently it was shown that MSL2 also interacts with roX2 close to the CLAMP binding domain (Muller et al., 2020; Valsecchi et al., 2021).

1.4.3 Mechanism of Function for the DCC

Dosage compensation is genetically encoded on the X chromosome through a 21 bp GA-rich motif termed the MSL recognition elements (MRE) (Alekseyenko et al., 2008; Straub et al., 2008), which is found in many of the about 309 so called high affinity sites (HAS) on the X chromosome (Straub et al., 2008). Initially, the complex is targeted to these sites by MSL2, which can bind the MREs through its C-terminal domain (Fauth et al., 2010; Villa et al., 2016; Zheng et al., 2014). HAS are scattered around the X chromosome and serve as nucleation sites for the dosage compensation process. From there, the DCC somehow “spreads” to close-by, active genes in a process that is not completely understood (Lionnet and Wu, 2021; Suter, 2020) and binds active genes trough MSL3 recognition of H3K36me3. It was also suggested that roX2-mediated condensates support the DCC targeting (Valsecchi et al., 2021). The propagation then leads to H4K16 acetylation of surrounding nucleosomes through MOF. The acetylation relieves chromatin-mediated repression of transcription in vitro and in vivo by disrupting histone-histone interactions (Akhtar and Becker, 2000; Zhang et al., 2017). The unfolding of chromatin by H4K16ac can be impressively seen at the level of polytene chromosomes (Bell et al., 2010). Decompaction by H4K16ac, however, is antagonized by the nucleosome remodeler NURF (Badenhorst et al., 2002; Corona et al., 2002).

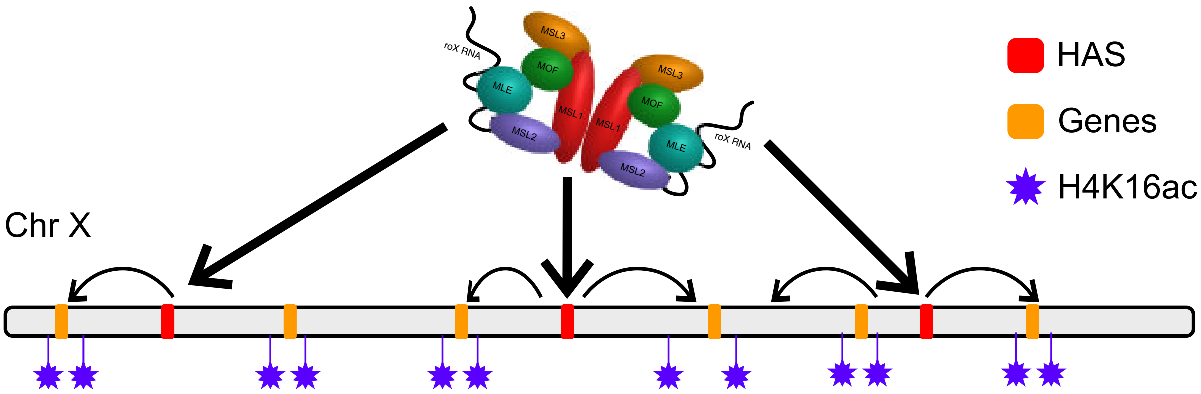


Figure 10: Spreading along the X chromosome by the DCC. The complex first binds to high affinity sites (HAS) on the x chromosomes through MSL2. From there it spreads to nearby genes on the same chromosome by binding H3K36me3 through MSL3. Finally, the complex acetylates H4K16 at active gene bodies through MOF.

Finally, H4K16 marks all active gene bodies on the X chromosome, increasing their transcriptional output roughly 2-fold and thereby to the levels present in females and to the other genes on the autosomes (Smith et al., 2001). This enhanced output is mostly facilitated through enhanced transcriptional elongation rather than increased transcriptional initiation (Kuroda et al., 2016; Larschan et al., 2011).

1.4.4 Targeting of the DCC through MSL2

According to the prevailing model, after the initial binding of MSL2 to the HAS on the X chromosome all further steps of the dosage compensation pathway are thought to involve sequence-unspecific spreading to nearby active genes and acetylation of their chromatin. High specificity of this first targeting step to the X chromosome is therefore imperative for this process to work properly. Inserting a HAS into autosomes is sufficient to recruit the whole complex to these ectopic sites (Alekseyenko et al., 2008; Dahlsveen et al., 2006). The MRE motifs that characterize HAS, however, are quite degenerate and are enriched only about 2-fold on the X chromosome. It is possible that multiple weak sites confer cooperative binding by the DCC (Gilfillan et al., 2007). Considering this, it is unclear how MSL2 distinguishes the MREs in a few hundred HAS on the X chromosome from the large excess of similar sequences at physiologically nonfunctional sites (Lucchesi and Kuroda, 2015). Hence, it was suggested that identifying chromatin and local sequence features may also contribute to the selection of functional over nonfunctional sites (Alekseyenko et al., 2012).

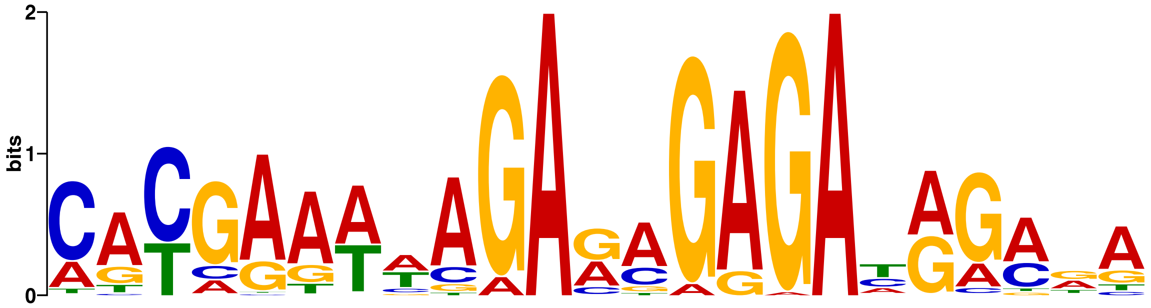


Figure 11: The PWM of PionX sites. PionX sites have a notable 5' extension of to the standard MRE motif. The height of each letter indicates the relative probability of base occurrence at each position, while the total height of all letters in a column combined indicates the information content of a position. In this case the PionX motif was derived from 56 known binding sites and determined by MEME.

Indeed, there is a subset of 56 HAS with a special sequence and DNA shape signature termed PionX (Pioneering sites on the X) for MSL2 (Villa et al., 2016). PionX sites are characterized through a notable 5′ extension of the known MRE motif and a particularly high DNA Roll shape feature at position +1 (explained in detail in 1.1.7). These sites are highly enriched on the X chromosome and are directly bound by MSL2 through its CXC-domain, explaining some of MSL2 binding specificity to the X. Previous research has also elucidated another interacting factor, CLAMP, which is discussed further in 1.5.1. Additionally, recent research has suggested a prominent role for roX RNA in targeting. In the presence of MLE, roX can incorporate into the MSL1-MSL2 complex and influence the chromatin binding specificity (Maenner et al., 2013). Also, it turns out MSL2 and roX can form phase-separated condensates and nascent roX RNA can concentrate MSL proteins around the loci of roX transcription (Kelley et al., 1999; Oh et al., 2003; Valsecchi et al., 2021). In absence of roX RNA, MSL2 is mistargeted to heterochromatic regions (Figueiredo et al., 2014).

1.5 Co-Factors of the DCC

1.5.1 The Chromatin Linked Adapter of MSL2 Proteins (CLAMP)

Due to its GA-richness, MREs resemble binding sites for known transcription factors that bind specifically to (GA)n sequences. Two such proteins that are relevant for this work are the CLAMP protein (Larschan et al., 2012) and the GAGA factor (GAF)(Wilkins and Lis, 1998). In vivo, both proteins bind thousands of GA-rich sites in both sexes. The PWM motifs of these two proteins are very similar to each other and the canonical MRE motif, yet GAF and CLAMP are rarely found at the same loci in vivo (Kaye et al., 2018). It was suggested that expansion of GA dinucleotides on the X chromosome, which led to higher density of CLAMP binding sites, was a driver for the evolution of dosage compensation.

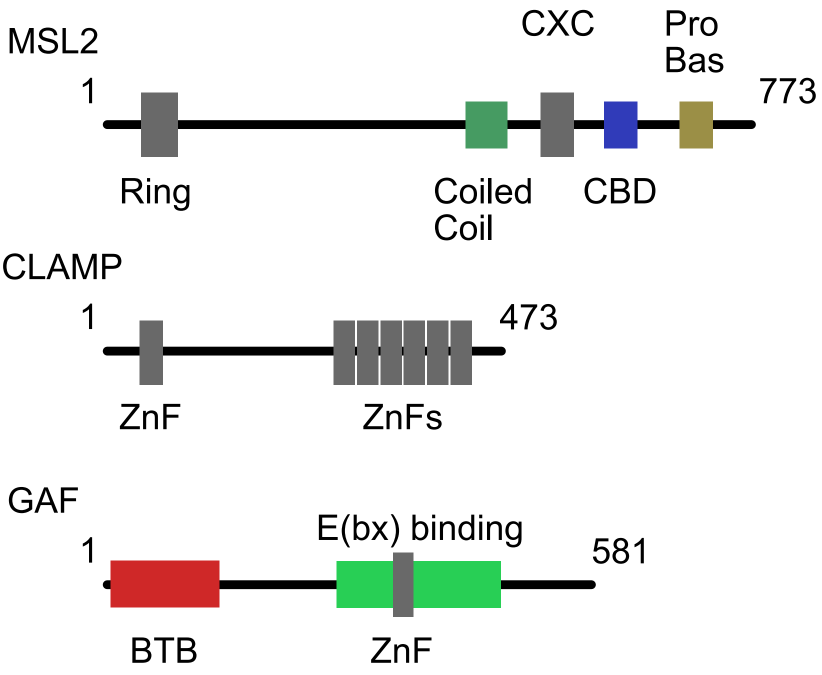


Figure 12: Domain distribution of GA binding proteins. Zn^{2+}-complexing domains are marked in grey. The MSL2 Ring domain is an E3 ubiquitin ligase, CXC is a DNA binding domain, CBD is the CLAMP binding domain and the Pro-Bas region is rich in prolines and basic residues. The GAF BTB domain (bric-a-brac) is a conserved protein interaction domain, while the E(bx) (enhancer of bithorax) binding domain is necessary for NURF301 binding.

CLAMP was first discovered using a genome-wide RNAi screen for factors required for DCC function (Larschan et al., 2012; Wang et al., 2013). CLAMP is a DNA binding protein with 7 C2H2 zinc fingers (ZnF)(Soruco et al., 2013a). Especially ZnF 4-7 have been shown to bind MREs (Soruco et al., 2013a). MSL2 and CLAMP physically interact at the N-terminal zinc finger domain of CLAMP and at a highly conserved region between 620-655 aa on MSL2, termed "CLAMP binding domain" (CBD). MSL2 and CLAMP cooperate to compete with nucleosome occupancy at HAS (Albig et al., 2019; Soruco et al., 2013a; Urban et al., 2017b; Urban et al., 2017c). Inactivation of either the CBD or CXC domain of MSL2 has only limited effects on DCC targeting, while double knockout has a severe effect, so this interaction appears to serve as redundant system assuring good MSL2 to X chromosomal binding (Tikhonova et al., 2019).

CLAMP is an essential protein that, while being implicated in MSL recruitment to chromatin, was also shown to increase global X chromosome accessibility in males through its recruitment of NURF (Urban et al., 2017a). It localizes to the majority of HAS in male cells, but only about 3% of all physiological CLAMP binding sites overlap with HAS (Albig et al., 2019; Urban et al., 2017b). It binds thousands of additional MREs throughout the genome and has functions outside dosage compensation such as histone locus regulation (Rieder et al., 2017; Urban et al., 2017c). CLAMP was also shown to interact with the Gypsy insulator complex and to promote the enhancer blocking and barrier forming function of the complex (Bag et al., 2019). This complicates the understanding of CLAMP's role in DCC recruitment. Nonetheless, CLAMP is essential for MSL2 targeting during early embryo development (Rieder et al., 2019) and in cells (Albig et al., 2019). Apparently, this very general GA-binding protein has been coopted by MSL2 for the specific task of stabilizing its association at GA-rich MREs. In vitro, DNA Immunoprecipitation (DIP) experiments showed that CLAMP does not enhance MSL2s X chromosomal specificity but only overall binding (Albig et al., 2019).

1.5.2 GAGA Factor (GAF)

GAF is the product of the trithorax-like gene (trl) and best known as a transcription factor that facilitates the chromatin association of other TFs in promoters, enhancers or polycomb response elements (Adkins et al., 2006; Fuda et al., 2015). It promotes chromatin accessibility and is essential for the zygotic genome activation during maternal-to-zygotic transition in early Drosophila embryos (Gaskill et al., 2021). It controls the expression of thousands of genes with specific GA-rich sequences in their regulatory regions by maintaining nucleosome-free regions at developmental promoters and recruits RNA Polymerase II (Fuda et al., 2015; Moshe and Kaplan, 2017; van Steensel et al., 2003). GAF minimally binds the pentamer GAGAG, though it also binds longer (GA)n stretches, as it forms multimers. This oligomerization enhances GAF specificity and affinity (Espinas et al., 1999). With respect to the DCC it was shown that mutants bearing a hypomorph allele of the trl gene show elevated levels of male-specific lethality and inappropriate binding of MSL2 to autosomes (Greenberg et al., 2004). GAF and CLAMP both recruit the nucleosome remodeler NURF (Judd et al., 2021; Urban et al., 2017a). This multitude of functions is attributed to the pioneering ability of GAF. It enables other factors to bind, through the removal of nucleosomes, which in turn is mediated by GAF interacting with chromatin remodelers (Chetverina et al., 2021; Tsukiyama et al., 1994; Wall et al., 1995). GAF and CLAMP rarely co-localize in chromatin suggesting discriminators in GA-rich sequences that provide exclusive selectivity (Kaye et al., 2018). However, the minimal recognition sequence for GAF should allow binding to many CLAMP sites (Wilkins and Lis, 1998). It seems that GA-rich sequences can be bound by several proteins with GA-binding potential and the occupancy is dynamically regulated. Through the expression of MSL2 in male cells, the GA-rich MREs are also subject to this cooperative and competitive regulatory system.

1.6 Aims

This thesis started with a specific predicament. Earlier experiments had shown that in vivo MSL2 enriches exclusively to the X chromosome. But in in vitro assays, where MSL2 is allowed to select binding sites from genomic DNA, it failed to enrich any higher than 50%, all the while still binding the same GA-rich sequence motif. So, the main issue addressed in this work is about identification of additional targeting determinants. How does MSL2 distinguish its functional targets from similar but nonfunctional ones? Because the DNA sequence alone does not instruct MSL2 properly, a combination of other factors must be involved. Also, the earlier DIP approaches lacked chromatin and thus did not include nucleosome competition. To address these questions, I reconstituted MSL2 binding in vitro, adding other protein components and measured how MSL2 targeting would change in response to these changing environments.

I employed genome-wide in vitro chromatin reconstitution to assess how MSL2, CLAMP and GAF interact with DNA targets in a complex and physiological chromatin environment. I aimed to elucidate, which binding properties these three GA-binding proteins show intrinsically, as well as which properties arise from cooperative and competitive interactions between them.

The selective targeting of the X chromosome is a vital requirement for balancing genome expression (Samata and Akhtar, 2018). This property of MSL2 can be instrumentalized to quantify in vitro assays. Since all functional binding sites must be on the X chromosome, any autosomal binding immediately exposes nonfunctional or “decoy” binding. This allows for a convenient readout for the effects of experimental manipulation in the cell-free genomics system. By calculating the enrichment of X-chromosomal binding events, one can assess the “quality” of the overall targeting.

The results are presented in three parts, which are: 1) The establishment and characterization of the reconstitution system, including the purification of all relevant factors for later studies, 2) the evaluation of the system, testing MSL2 and CLAMP individually and comparing these first results with earlier findings to establish the proper in vitro settings, and 3) the utilization of said system to demonstrate novel concepts of protein cooperation and competition, for MSL2 targeting in particular and transcription factor binding in general.

2. Materials and Methods

2.1 Materials

2.1.1 Antibodies

NAME	Species	Type/Name	Application	Source
α - MSL1	Rabbit	Polyclonal	ChIP 1 µl	E. Schulze
α - MSL2 1A8	Rabbit	Monoclonal	ChIP 1 ml	(Albig et al., 2019)
α - MSL2	Guinea pig	Polyclonal	WB 1:1000	C. Regnard (Pineda)
α - MLE	Rat	Monoclonal	WB 1:10000	E. Kremmer (Helmholtz)
α - CLAMP 2C7	Rabbit	Monoclonal	ChIP 1 ml	(Albig et al., 2019)
α - GAF	Rabbit	Polyclonal	ChIP 1 µl	(Strutt et al., 1997)
α – NURF-301	Rabbit	Polyclonal	ChIP 1 µl	Paul Badenhorst
α – H2A.V	Rabbit	Polyclonal	ChIP 1 µl	(Anton Eberharter)
α – γH2A.V	Mouse	Monoclonal	ChIP 1 µl	(Rockland)
α – H3 ab1791	Rabbit	Polyclonal	ChIP 1 µl	Abcam
IRDye 800CW	Goat	Monoclonal	WB 1:10000	(LI-COR Biosciences)
IRDye 680RD	Goat	Monoclonal	WB 1:10000	(LI-COR Biosciences)

2.1.2 Buffers

Name	Recipe
Agarose gels	0.5x TAE buffer 2% Agarose 1 µg/ml Ethidiumbromide
Buffer C	50 mM HEPES pH 7.6 1 M KCl 1 mM $MgCl_2$ 5% (v/v) Glycerol 0.05% NP-40 50 µM $ZnCl_2$ 375 mM L-Arginine (Leibly et al., 2012)
Coomassie fixing	50% Ethanol 10% Acetic acid
Coomassie staining	5% Ethanol 7.5% Acetic Acid 0.0025% Coomassie Blue (w/v)
DNA binding buffer	100 mM KCl 2 mM $MgCl_2$ 2 mM Tris-HCl pH 7.5 10% Glycerol 10 µM $ZnCl_2$
Embryo-Wash	0.7% NaCl 0.04% Triton X-100
EX10	10 mM Hepes-NaOH pH 7.6 10 mM KCl 1.5 mM $MgCl_2$ 10% (v/v) Glycerol 1x PIC
EX50	10 mM Hepes-NaOH pH 7.6 50 mM NaCl 1.5 mM $MgCl_2$ 10% (v/v) Glycerol

	1x PIC 10 µM $ZnCl_2$
HEMG 2.1 AS	25 mM Hepes-KOH pH 7.6 12.5 mM $MgCl_2$ 0.1 mM EDTA pH 8.0 10% (v/v) Glycerol 2.1 M $(NH_4)_2SO_4$ 1 mM DTT 0.2 mM PMSF
HEMG 0.1 AS	25 mM Hepes-KOH pH 7.6 12.5 mM $MgCl_2$ 0.1 mM EDTA pH 8,0 10% (v/v) Glycerol 100 mM $(NH_4)_2SO_4$ 1 mM DTT 0.2 mM PMSF
HEMG 100	25 mM Hepes-KOH pH 7.6 100 mM NaCl 12.5 mM $MgCl_2$ 0.1 mM EDTA pH 8.0 10% (v/v) Glycerol 1 mM DTT 0.2 mM PMSF
HEMG 40	25 mM Hepes-KOH pH 7.6 40 mM NaCl 12.5 mM $MgCl_2$ 0.1 mM EDTA pH 8,0 10% (v/v) Glycerol 1 mM DTT 0.2 mM PMSF
HEMG 1000	25 mM Hepes-KOH pH 7.6 1 M NaCl 12.5 mM $MgCl_2$ 0.1 mM EDTA pH 8,0 10% (v/v) Glycerol 1 mM DTT 0.2 mM PMSF

High-Salt Buffer	10 mM Tris-HCl pH 7.6 2 M NaCl 1 mM EDTA pH 8.0 0.05% Igepal CA-630 0.1% β-mercaptoethanol
Low-Salt Buffer	10 mM Tris-HCl pH 7.6 50 mM NaCl 1 mM EDTA 0.05% Igepal CA-630 0.01% β-Mercaptoethanol
Laemmli buffer 5x	250 mM Tris-HCl pH 6.8 50% Glycerol (v/v) 10% SDS (w/v) 0.05% Bromophenol Blue (w/v) 0.5 M DTT
McNAP 10x	30 mM $MgCl_2$ 10 mM DTT 300 mM Creatine phosphate 30 mM ATP 10 µg/ml Creatine phosphate kinase (CPK in 100 mM imidazole pH 6.6)
MNase solution	EX50 5 mM $CaCl_2$ 12 U/ml MNase (in EX50)
MSL lysis buffer	300 mM KCl 50 mM Hepes/KOH pH 7.6 5% glycerol 0.05% NP-40 1 mM $MgCl_2$ 50 µM $ZnCl_2$ 1 mM DTT 1x PIC
MSL wash buffer	MSL lysis buffer but 1 M KCl and 1% NP-40

MSL elution Buffer	MSL lysis buffer but 100 mM KCl
NU-1	15 mM HEPES-KOH pH 7.6 10 mM KCl 5 mM $MgCl_2$ 0.5 mM EGTA 0.1 mM EDTA 350 mM Sucrose 1 mM DTT 1x PIC 1 mM $Na_2S_2O_5$
NU-2	15 mM HEPES-KOH pH 7.6 110 mM KCl 5 mM $MgCl_2$ 0.1 mM EDTA 1 mM DTT 1x PIC 1 mM $Na_2S_2O_5$
NXI buffer	15 mM HEPES-KOH pH 7.5 10 mM KCl 5 mM $MgCl_2$ 0.1 mM EDTA 0.5 mM EGTA 350 mM Sucrose 1 mM DTT 0.2 mM PMSF 1 mM $Na_2S_2O_5$
Orange-G loading dye 6x	60% glycerol (v/v) 40% TE buffer 2 mg/ml Orange G
PBS (T) 10x	1.4 M NaCl 27 mM KCl 100 mM Na_2HPO_4 18 mM KH_2PO_4 (0.1% Tween-20)
RIPA buffer	25 mM Hepes-NaOH pH 7.6 150 mM NaCl

	1% Triton-X-100 0.1% SDS 1 mM EDTA 0.1% Na-deoxycholate 1 mM PMSF 1x PIC
RNAse A	10 mg/ml RNAse A 10 mM Tris-HCl pH 7.5 15 mM NaCl heated to 100°C for 15 min
SDS running buffer	25 mM Tris pH 8.3 192 mM Glycine 0.1% SDS
Suc buffer	15 mM HEPES-KOH pH 7.5 10 mM KCl 5 mM $MGCl_2$ 0.05 mM EDTA 0.25 mM EGTA 1.2% (v/v) Sucrose 1 mM DTT 0.1 mM PMSF
TBS(T) 10x	250 mM Tris-HCl pH 8 30 mM KCl 1400 mM NaCl (1% Tween-20)
TE buffer	10 mM Tris-HCl pH 8 1 mM EDTA
WB transfer buffer	25 mM Tris 192 mM Glycine

2.1.3 Cell Lines

Name	Species	Application	Source
BG3-c2	D. melanogaster	Genomic DNA purification	DGRC
SF21	D. melanogaster	Baculo virus infection, protein purifications	DGRC

2.1.4 Chemicals

Acetic Acid (CLN); Adenosine triphosphate (ATP, Sigma); Agarose (Bio & Sell); Ammonium acetate (Roth); Ammonium sulfate (NH_4SO_2, Merck Millipore); AMPure XP DNA beads (Beckman Coulter); Ampicillin (Roth); β-Mercaptoethanol (Serva);Bovine serum albumin (BSA, Sigma); Chloroform (NeoLab); Creatine Phosphate (Roche); cOmplete Protease inhibitor (PIC, absource); Coomassie Blue G250 (Serva); DAPI (Invitrogen); DMSO (Sigma); DTT (Roth); EDTA (Diagonal); EGTA (Carl Roth); Ethanol (VWR); Ethidium Bromide (VWR); Fetal Calf Serum (FCS, Sigma); Glycerol (VWR); Glycine (VWR); Hepes (Serva); Igepal (Sigma);KCl (VWR); L-Arginine (Roth); 2-Mercaptoethanol (Sigma); Methanol (CLN); $MgCl_2$ (VWR); NaCl (Serva); Na-deoxycholate (Sigma);$Na_2S_2O_5$ (VWR); $(NH_4)_2SO_4$ (MP Biomedicals); NP-40 (Sigma); Nipagin (Sigma-Aldrich); Penicillin/Streptavidin (Life Technologies); 37% Formaldehyde (Merck Millipore); Phenylmethylsulfonylfluoride (PMSF, Genaxxon); Phenol:Chloroform:-Isoamylalcohol (Invitrogen); Polyethylene imine (Sigma-Aldrich) ; 2-Propanol (Sigma); Propionic acid (Sigma); Sodium acetate (Sigma); Sodium azide (Merck); Sodium deoxycholate (Sigma); Sodium dodecyl sulfate (SDS, Serva); 6-14% Sodium hypochlorite (Merck Millipore); Sucrose (VWR);Sugar (Südzucker); Tris (Diagonal); Triton X-100 (Sigma); Tween-20 (Sigma); Nuclease-free water (Invitrogen); $ZnCl_2$ (Merck); Zuckerrübensirup (Bauk)

2.1.5 Consumables and Instruments

Agar-Agar (Die Gewürzmühle Brecht)
Amicon Ultra-4 Centrifugal filter (Merck Millipore)
AMPure XP DNA beads (Beckman Coulter)
Apple juice (Discounter)
Avanti JXN-26 Centrifuge (Beckman Coulter)
Bioanalyzer (Agilent)
25, 75, and 170 cm^2 Cell Culture Flasks (Greiner)
Cellulose (Arndt)
12 mm round Coverslip (Paul Marienfeld)
Dialysis cups Slide-A-Lyzer, MWCO 3500 (Thermo Fisher)
Dry yeast, Fermipan rot (Hobbybäcker)
1.5- and 2- mL Eppendorf tubes
15- and 50- mL Falcon tubes
Homogenplus homogenizer (Schuett-Biotec),
Miracloth (Sigma Aldrich)
NanoDrop (Thermo Fisher Scientific)
27G needle (B. Braun)
Open-Top Thinwall Ultra-Clear Tube, 14 x 89 mm (Beckman Coulter)
Optima XPN-80 Ultracentrifuge (Beckman Coulter)
Peristaltic pump (Minipulse evolution, Gilson, mode 8.4 rpm).

Phenyl Sepharose 6 Fast Flow (Sigma Aldrich)
6-, 12-well plate (Sarstedt)
Qubit fluorometer (Thermo Fisher Scientific)
Schneider’s Drosophila medium (life technologies)
Sepharose protein-A and -G beads (Helmholtz Centre Munich, E. Kremmer)
SiR-DNA (Spirochrome)
Styrofoam dishes (Margret Lutz)
Syringes (Injekt)
Table-top centrifuge (Eppendorf)
Thermocycler (Eppendorf)
Yamato LH-21 homogenizer (Triad Scientific)
Yeast extract (BD Biosciences)
XK16/20 column (Cytiva)

2.1.6 Kits, Enzymes, Markers

Name	Source
Asc I	NEB
100 bp and 1 kb DNA markers	NEB
Creatine phosphate kinase	Roche
DNeasy Blood & Tissue kit	Qiagen
DNA1000/ HS bioanalyzer kit	Agilent
ECL advance western blotting detection kit	VWR
Fast Sybr Green master mix	Applied Biosciences
Maxtract High Density Kit	Qiagen
Micrococcal nucleoase	Sigma
Nebnext Ultra II DNA library	New England Biolabs
Not I	NEB
Orange G	Sigma
Proteinase K	Qiagen
Qubit ds DNA HS assay kit	Life Technologies

Restriction enzymes	NEB
Rnase A	Sigma
Triple colour protein standard iii	Sigma

2.1.7 Oligonucleotides

Name	Sequence forward	Sequence reverse
Transcription elongation factor SPT4	GCTCCGATTCATAAGCCCAG	GCCTCTTTCGGAGCAGCTTT
Highwire (HIW)	TCATCAGATTGGCACTGCAC	AACCGTGTTCTTCCATCTCG
Tomosyn (TOM)	CGGGCAATAGTCTGCAATG	TTGCTTGGTTGTGTGCGTAT
Not I	GAAACCCTAACACAG-GATGC	TGGTGGTGGTGAAGATGATG
Asc I	CAATT-GTTTAGCTCAATTTAAGGCG	GGTAGTGGAACAG-CAAGGTGAT

2.2 Large-Scale Drosophila Populations

2.2.1 Fly Populations

Oregon-R (wild-type) flies were used for all experiments and maintained at 25°C and appropriate humidity according to standard protocols.

2.2.2 Agar and Yeast

Recipe for about 200 Apple juice agar plates: In a 50 L heated stirrer, 1460 g agar were dissolved in 32 L deionized water at about 100°C, then 14 L apple juice and 2 L sugar beet sirup were slowly added while stirring. The agar was allowed to sit until cooled to 70°C, then 1120 mL 10% (w/v) Nipagin (dissolved in ethanol) was added. Immediately, the agar juice was poured into the plates [225 cm x 175 cm x 25 cm size (Margret Lutz GmbH & Co. KG)] to about 1 cm height and left to solidify. Plates were stored in clean plastic bags at 4°C. For feeding, 500 g dry yeast, 750 mL H_2O and 4.2 mL propionic acid (Sigma) were stirred to yield a moist paste and stored at 4°C. About 10 ml of this paste was added to each plate directly before adding them to the fly cages.

2.2.3 Embryo Boxes

Boxes were prepared as previously described (Harpprecht, 2018). Embryos from overnight plates were collected, washed in 70% EtOH and resuspended in EW buffer. About 5 ml of embryo solution per cage were transferred to small Whatman papers to dry. Cellulose tissues were distributed to plastic boxes of about 30 cm x 30 cm x 10 cm and soaked with 425 mL - 450 mL of an embryo nutrition suspension composed of 77 g inactivated yeast, 50 g sugar, 12 mL 10% (w/v) Nipagin (dissolved in ethanol), 4.8 mL ortho-phosphoric acid, 0.6 mL propionic acid and 430 mL water. The Whatman papers with embryos were transferred onto the soaked cellulose tissues within the boxes, after which the boxes were closed with finely latticed lids to allow air circulation. The boxes were then kept at 25°C and appropriate humidity for 10 days until the embryos reached the adult stage. After hatching, flies in the boxes were transferred in to collection cages and kept at 25°C and appropriate humidity. Agar plates stocked with yeast paste were added and changed regularly. On day 4 (day 14 of the cycle) the overnight plates were collected and used for the next round of embryo boxes.

2.2.4 Drosophila Embryo Extract (DREX) Preparations

DREX preparations were performed as described before (Eggers and Becker, 2021). DREX was prepared from preblastoderm Oregon-R fly embryos collected repeatedly in 90 minute intervals (Becker and Wu, 1992). 50 ml of settled embryos were dechorionated in 200 ml embryo wash buffer (EW) and 60 ml 13% sodium hypochlorite (VWR) for 3 min at room temperature (RT) while stirring. Embryos were rinsed for 5 min on a sieve with cold water and transferred into a glass cylinder with EW. Settled embryos were washed first in 0.7% NaCl and then in EX10. Embryos were settled in a homogenplus homogenizer (Schuett-Biotec), the supernatant was decanted and the embryos were homogenized with one stroke at 3000 rpm and 10 strokes at 1500 rpm. The $MgCl_2$ concentration of the homogenate was adjusted to 5 mM (final concentration) and centrifuged for 15 min at 27,000 g at 4°C. The white lipid layer was discarded and the supernatant was centrifuged for 2 h at 245,000 g at 4°C. The clear extract was collected with a syringe, leaving the lipid layer and pellet behind. Protein concentrations were measured using NanoDrop. Extracts were stored in 200 µl aliquots - 80°C after shock frosting in liquid N_2. Extracts were only thawed once before use. EDTA was excluded from all steps of the purification to avoid chelation of Zn^{2+}.

2.2.5 Transcriptionally Active Extract (TRAX) Preparations

TRAX was prepared from Oregon-R wildtype fly embryos collected every 12 h up to a total amount of 100 g. Collected embryos were kept at 4°C during the further collections to halt development. Then embryos were dechorionated in 200 ml embryo wash buffer (EW) and 60 ml 13% sodium hypochlorite (VWR) for 3 min at RT while stirring. Embryos were rinsed for 5 min on a sieve with cold water and dried using a towel mesh. Then embryos were dispersed in 3 ml/g of buffer NU-1 and disrupted by passing them 6 times through a Yamato LH-21 homogenizer at 1000 rpm. Homogenate was filtered through Miracloth and then washed with 2 ml/g of NU-1. Homogenate was then centrifuged at 18,000 g for 15 min at 4°C. Supernatant was decanted and pellet was resuspended in 1 ml/g NU-1. Then 1/10 Volume of 4 M ammonium sulfate was added and samples were incubated for 20 minutes. Then samples were pelleted at 125,000 g for 60 minutes. The pellet was left behind and the supernatant was further precipitated by adding 0.3 g/ml finely ground ammonium sulfate over a 5-minute period. Then samples were centrifuged again for 30 minutes at 27000 g at 4°C. The supernatant from this step was used in the H1 purification protocol described later. For the TRAX preparations the supernatant was aspirated and the pellet was resuspended in 200 µl of HEMG-40 per gram of embryos. Then samples were dialyzed against HEMG40 until the conductivity was equal to HEMG100. Precipitated proteins were spun out by centrifugation at 10,000 g and samples were flash frozen in liquid nitrogen and stored at -80°C.

2.3 Cell Biology

2.3.1 Cell Maintenance

Cells were cultured at 26°C in Schneider's Drosophila Medium (GIBCO) with 10% fetal calf serum (FCS), penicillin-streptomycin and 10 mg/ml human insulin and regularly tested for mycoplasma.

2.3.2 Genomic DNA Purification

Genomic DNA (gDNA) was purified from male BG3-c2 cells (Drosophila Genomics Resource Center) because they show the best male ploidity with the fewest deletions and duplications in their genome (Lee et al., 2014). The DNA of 10^7 cells was purified using the Blood & Cell Culture DNA Midi Kit (Qiagen) following the supplier's protocol. The resulting DNA was dissolved in EDTA-free 10 mM Tris-NaCl pH 8. Concentrations were determined using Qubit (Thermo Fisher).

2.3.3 Baculovirus Infections

Sf21 cell cultures at 10^6 cells/ml (2.5 x10^8 cells) were infected 1:1000 (v/v) with baculovirus, expressing the respective FLAG-tagged proteins as described in (Fauth et al., 2010). After

72 h, cells were harvested and washed once in PBS, frozen in liquid nitrogen and stored at -80°C.

2.4 Protein purification protocols

2.4.1 Histone – Octamer Purification

The protocol was adapted from (Krietenstein et al., 2012; Simon and Felsenfeld, 1979). At first Oregon-R fly embryos were collected every 12 h up to a total amount of 100 grams. Collected embryos were kept at 4°C during the further collections to halt development. Then embryos were dechorionated in 200 ml embryo wash buffer (EW) and 60 ml 13% sodium hypochlorite (VWR) for 3 min at RT while stirring. Embryos were rinsed for 5 min on a sieve with cold water. Embryos were resuspended with 40 ml NXI buffer per 50 g of embryos and disrupted by passing them 6 times through a Yamato LH-21 homogenizer at 1000 rpm. Samples were filtered through a Miracloth and centrifuged at 7000 g for 15 minutes at 4°C. This results in three phases, a solid pellet, a brownish jelly and a liquid supernatant with some lipids on top. The supernatant was decanted and the middle layer was resuspended carefully with 50 ml SUC buffer per 100 grams of embryos. This sample was transferred to a fresh tube and centrifuged at 7000 g for 15 min at 4°C. Supernatant was decanted and the pellet was washed once again in SUC buffer. After decanting the supernatant, the pellet was resuspended in 30 ml SUC buffer per 50 grams of embryos and dounced 20 times with a B-pestle homogenizer. Then 90 µl 1 M $CaCL_2$ per 50 g of embryos are added and the solution is warmed to 26°C in a water bath. Samples are digested with 125 µl MNase solution (50U/µl) for 10 minutes at 26°C, then stopped with 500 µl of 0.5 M EDTA pH 8. Samples are centrifuged at 10000 g for 15 min at 4°C. Supernatant is discarded and pellet is resuspended in 6 ml TE pH 7.6; 1 mM DTT; 0.2 mM PMSF, 1x PIC per 50 g of embryos. Then samples are centrifuged again at 15000 g for 30 min at 4°C and the supernatant is collected. Then salt concentration is adjusted to 0.63 M KCl by adding 2 M KCl, 0.1 M KH_2PO_4 buffer. Samples are centrifuged again at 15000 g for 15 min at 4°C and supernatant is filtered at 0.45 µm and again with 0.22 µm.

The samples are added to a hydroxylapatite column prewashed with 2 column volumes 0.63 M KCl, 0.1 M KH_2PO_4 buffer. Samples are washed with 23 column volumes of 0.63 M KCl, 0.1 M KH_2PO_4 buffer, then the buffer is changed to 2 M KCl, 0.1 M KH_2PO_4 buffer to elute the samples. Fractions were collected and 20 µl of selected fractions were denatured with 5 µl of 5x Laemmli buffer on 95°C for 5 min. Samples were then run on a 12% SDS-Page at 200 V for 90 min. Samples were fixed with Coomassie fixing solution and stained with Coomassie staining solution for 15 min, then destained with water. Samples detecting histone octamers were then pooled for further purification. Then samples concentrated on a 10 kDa 15 ml Amicon filter by centrifugation at 5000 g at 4°C for 10 min. Samples were adjusted to a final concentration of 40% glycerol, 5mM DTT and 1x PIC and stored at -20°C. Protein concentrations were determined via SDS–PAGE and Coomassie staining using BSA (NewEngland Biolabs) as a standard.

2.4.2 Histone H1 Purification

The protocol was adapted from (Croston et al., 1991). Phenylsepharose columns were prepared as follows. A XK 16/20 column as washed with 20% ethanol and filled with “Phenyl SepharoseTM 6 Fast Flow” by pouring. It is important not to trap any gases. After closing the adapter, the column was connected to an ÄKTA pure system and packed by 3 ml/min flow until media height remained constant.

Samples were prepared as described in 2.2.5 and the supernatant after the ammonium sulfate precipitation was filtered through 0.45 µm filter and the packed phenylsepharose column was equilibrated with the HEMG 2.1 AS buffer. Sample supernatant was loaded into the sample pump at 1 ml/min. Samples were eluted under a gradient from 2.1 ammonium sulfate (AS) to 0.1 AS in HEMG. Fractions were collected and 20 µl of selected fractions were denatured with 5 µl of 5x Laemmli buffer on 95°C for 5 min. Samples were then separated by 12% SDS-PAGE at 200 V for 90 min. Samples were fixed with Coomassie fixing solution and stained with Coomassie staining solution for 15 min, then destained with water. Samples detecting H1 were then pooled for further purification.

Next, samples were centrifuged at 40,000 g for 30 min at 4°C and the supernatant filtered with 0.2 µM filters. A “Mono S 5/50 GL” column was loaded and equilibrated with HEMG100. Flowrate was set to 2 ml/min and samples were loaded. Then the column was washed with HEMG100 buffer for 10 column volumes. Samples were eluted on a gradient from 0-100% HEMG100 to HEMG1000 and collected in fractions. Again, 20 µl of selected fractions were denatured with 5 µl of 5x Laemmli buffer on 95°C for 5 min. Samples were then run on a 12% SDS-Page at 200 V for 90 min. Samples were fixed with Coomassie fixing solution and Stained with Coomassie staining solution for 15 min, then destained with water. Samples detecting H1 were then pooled and concentrated on a 10 kDa 15 ml Amicon filter by centrifugation at 4000 g at 4°C for 10 min. Finally, samples were transferred to a new tube and concentrations were determined by Bratford assay. Samples were then flash frozen in liquid nitrogen and stored at -80°C.

2.4.3 MSL2 Purification

Sf21 cell pellets previously infected with baculoviruses as described before (Fauth et al., 2010), were rapidly thawed and resuspended in 25 ml ice-cold MSL Lysis buffer per cell pellet (2.5 x10^8 cells). After 15 min incubation on ice, the suspension was sonicated (5x10 sec pulses, 20 sec break, 20% amplitude, Branson digital sonifier model 250-D) and centrifuged for 45 min at 30,000 g at 4°C. The soluble protein fraction was incubated with Lysis buffer equilibrated Agarose M2 FLAG beads for 3 h at 4°C on a rotating wheel. 0.5 ml beads were used per 2.5 x10^8 cells. The beads were washed twice with 10 ml ice-cold Lysis buffer, twice with 10 ml Wash buffer and twice with 10 ml MSL Elution buffer. The FLAG-tagged MSL proteins were eluted for 3 h at 4°C on a rotating wheel in the presence of 0.5 mg/ml FLAG-Peptide (Sigma) in 1 ml Elution buffer. Purified proteins were then rapidly frozen in liquid nitrogen and finally stored at -80°C. Protein concentrations were determined via SDS–PAGE and Coomassie staining using BSA (NewEngland Biolabs) as a standard. Cloning for the MSL2 expression construct is described by (Fauth et al., 2010).

2.4.4 CLAMP Purification

Sf21 cell pellets were rapidly thawed and resuspended in 1 mL Buffer C per 10 mL of culture [according to (Leibly et al., 2012)] supplemented with 0.5 mM TCEP and 1x PIC. After 15 min incubation on ice, the suspension was sonicated (5x10 sec pulses, 20 sec break, 20% amplitude, Branson digital sonifier model 250-D). The extract was adjusted with Buffer C containing PI to 2 mL per 10 mL of culture and supplemented with 0.1% (v/v) polyethyleneimine by adding 2% (v/v) polyethyleneimine (neutralized with HCl to pH 7.0) drop-by-drop while string in an ice bath [according to (Patel et al., 2016)] and then centrifuged for 45 min at 30,000 g at 4°C. The soluble protein fraction was incubated with Buffer C equilibrated FLAG beads (Anti-FLAG M2 Agarose, Sigma) for 3 h at 4°C on a rotating wheel. 0.5 ml beads were used per 2.5 x10^8 cells. Beads were pelleted at 4°C for 5 min at 500 g and supernatant was removed. Beads were washed 5 times with 20 bed volumes of Buffer C. The FLAG-tagged CLAMP proteins were eluted for 3 h at 4°C on a rotating wheel in the presence of 0.5 mg/ml FLAG-Peptide (Sigma) in 1 ml Buffer C containing 1x PIC. Purified proteins were then rapidly frozen in liquid nitrogen and finally stored at -80°C. Protein concentrations were determined via SDS–PAGE and Coomassie staining using BSA (NewEngland Biolabs) as a standard. Cloning for the CLAMP construct is described by (Albig et al., 2019).

2.4.5 GAF Purification

GAF was purified identically as described in the MSL2 purifications, with the difference that Cells were infected with the FLAG-GAF expression construct instead of the MSL2. Cloning for the GAF expression construct is described in (Wall et al., 1995) and was re-cloned for this work by Silke Krause.

2.5 Molecular Biology Methods

2.5.1 General Molecular Biology Methods

General molecular biology methods are done according to standard protocols.

2.5.2 Salt Gradient Dialysis-assisted Chromatin Assembly

The protocol was adapted from (Harpprecht, 2018). To assemble nucleosomes 10 µg DNA, around 10 µg octamer (optimal octamer amounts were determined by titration), 20 µg BSA, 0.1% Igepal CA-630, 10 mM Tris at pH 7.6, 2 M NaCl, 1 mM EDTA in 100 µl total volume were transferred into dialysis cups and placed in 300 mL high salt buffer. Salt concentration was decreased constantly at RT by pumping 3 L of low salt buffer into the 300 mL high salt buffer using a peristaltic pump (Minipulse evolution, Gilson, mode 8.4 rpm). After the gradient, the dialysis cups were dialyzed 2 h at RT against low salt buffer. Quality of nucleosome assembly was assessed by limited MNase digestion.

2.5.3 DREX-Assisted Chromatin Assembly

Chromatin assemblies were performed as described before (Eggers and Becker, 2021). 1 µg of genomic DNA was assembled into chromatin by adding 15 µl McNAP buffer, 100 µl DREX extract and up to 150 µl total amount EX50 buffer. Exact amounts of extract necessary were determined empirically for each batch. Assembly took place at 26°C for 4 h at 300 rpm on a shaking heat block.

2.5.4 Micrococcal Nuclease Digestion Protocol

Digestions were performed as described before (Eggers and Becker, 2021). 1 µg of DNA assembled into chromatin in 150 µl solution was digested with MNase by adding 200 µl MNase digestion solution. At times 15, 30 and 120 sec, 110 µl were transferred to tubes containing to 40 µl 100 mM EDTA solution each to stop the digest. 2 µl glycogen (10 mg/ml) and 150 µl 7.5 M ammonium acetate were added and samples were mixed. Then, 880 µl of 100% ethanol was added and samples were vortexed vigorously and cooled at -20°C for 10 min. After centrifugation at 21,000 g for 15 min at 4°C, the supernatant was removed and pellets were washed with ice-cold 70% ethanol. After pelleting the DNA again at 21,000 g for 5 min at 4°C, it was dissolved in 8 µl 10 mM TE buffer and 2 µl Orange-G loading dye. Samples were separated on a 2% agarose gel pre-stained with ethidium bromide and imaged using the Quantum ST-4 from PeqLab.

2.5.5 DNA Immunoprecipitation

Dip-Seq experiments were performed as in (Gossett and Lieb, 2008) with some modifications. 400 ng of genomic DNA (gDNA) were added to 80 nM of MSL2-FLAG at 26°C for 30 min in 100 µl of DIP binding buffer. 10% of the reactions was taken as the input and subjected to quantitative PCR and/or deep sequencing. DNA-protein complexes were immunoprecipitated using 15 µl of FLAG bead slurry (M2, SIGMA) for 15 min at RT and washed twice with 100 µl of DIP binding buffer to eliminate unbound DNA. After proteinase K digestion (0.5 mg/ml, 1 h at 56°C), DNA was purified with the GenElute kit (SIGMA) and subjected to quantitative PCR and/or deep sequencing.

2.5.6 Chromatin in vitro Immunoprecipitation

Chromatin immunoprecipitations were performed as described before (Eggers and Becker, 2021). Recombinant Proteins were added to 1 µg chromatin, assembled as described in 2.5.3, and were allowed to bind for 1 h. Samples were crosslinked by 0.1% formaldehyde for 5 min and then quenched by 125 mM glycine for 10 min. Samples were partially digested by MNase as described under 'MNase digestions' for 2 min. After adding 1x RIPA buffer up to 500 µl samples were precleared on a rotating wheel with 20 µl protein AG beads per 1 µg chromatin for 1 h at 4°C.

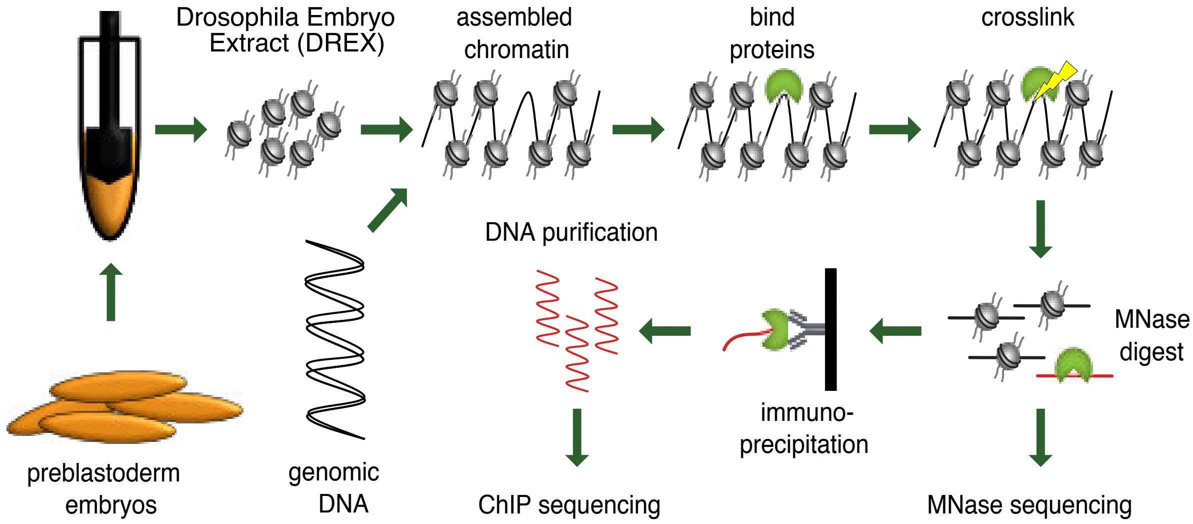


Figure 13: Schematic depiction of the ChIP-seq and MNase-seq workflow.

For immunoprecipitations with monoclonal antibody 20 µl beads per sample were bound to monoclonal antibodies by adding 1 ml of culture supernatant and rotating at 4°C for 3 h. Beads were pelleted at 1000 g for 1 min and the supernatant discarded. Antibody-coated beads were washed once with 1x RIPA buffer. The precleared samples were pelleted at 1000 g for 1 min and supernatant was transferred to the antibody-bound beads. Binding was done overnight at 4°C on a rotating wheel.

For immunoprecipitations with polyclonal antibody 1 µl of purified antibody was added to the precleared sample supernatant and let to bind overnight at 4°C on a rotating wheel. Then samples were bound to freshly washed protein AG beads for 3 h.

Both kinds of antibody bound samples were washed 4 times for 5 min with 1 ml of 1x RIPA buffer per sample (1 µg chromatin on 20 µl beads). The beads were suspended in 100 µl 1x TE buffer and de-crosslinked overnight at 65°C while shaking. Samples were then digested with 10 µg RNAseA for 30 min at 37°C and 100 µg proteinase K at 56°C for 1h. Beads were pelleted at 1000 g for 1 min and supernatant was transferred to a fresh tube for purification.

2.5.7 DNA Purification

DNA purifications were performed as described before (Eggers and Becker, 2021). DNA was purified by two extractions with Phenol:Chloroform:Isoamyl-alcohol (25:24:1, Sigma Aldrich) and precipitated by adding it to 2 µl of glycogen, 0.1 x volume 3 M sodium acetate and 2.5 x volume 100% ethanol, cooling at -20°C for 15 min and pelleting in a tabletop centrifuge. The DNA was washed once with 70% ethanol and dissolved in EDTA-free 10 mM Tris/NaCl, pH 8. Concentrations were determined using Qubit (Thermo Fisher).

2.5.8 Library Preparations for NGS

Next generation sequencing libraries were prepared using NEB Next Ultra II DNA Library (New England Biolabs) according to manufacturer's instructions and sequenced by the Laboratory for Functional Genome Analysis (LAFUGA), Gene Center Munich, Germany using an Illumina HiSeq1500 sequencer. About 25 million paired-end reads were sequenced per sample for ChIP samples and 60 million paired-end reads for MNase-sequencing samples. Base calling was performed by Illumina's RTA software, version 1.18.66.3.

2.5.9 Fluorescence Microscopy

Fluorescence microscopy was performed as described before (Eggers and Becker, 2021). A standard chromatin assembly reaction was allowed to proceed for 4 hr. The DNA was stained with 1 µM final concentration SiR-DNA (Spiro-Chrome) for 15 minutes at RT. Omitting either DNA (extract only) or DREX (DNA only) served as controls. Samples were placed in sealed sample chambers made by punching a hole into a double-sided sticky tape, which was in turn taped onto a glass slide and sealed with a coverslip. Widefield fluorescence microscopy was performed at the Core Facility Bioimaging of the Biomedical Center with an inverted Leica DMi8 microscope, equipped with a SPECTRA X light engine from Lumencor and a Leica DFC365 FX CCD camera. Images were acquired with a 63x/1.4 NA oil immersion objective; image pixel size was 102 nm. SiR-DNA was excited with 13% power of the SPECTRA X light engine red LED with an effective excitation range of 625-650 nm. The emitted signal was detected with a quad band filter cube with the relevant emission band 670–770 nm. The exposure time was set to 200 ms. The protocol was adapted from (Eggers and Becker, 2021)

2.5.10 ATP Depletion of Extracts for ChIP

To test the requirement for ATP for factor binding, I used the ATP depletion protocol as described before (Tatei et al., 1989). In short, a standard chromatin assembly was performed for 4 hours. Then ATP was depleted by addition of 60 mM glucose (1,5 x of combined ATP and creatine phosphate concentration in the samples) and 6 U of hexokinase for 30 minutes. After depletion external factors were added and the standard ChIP protocol proceeded.

2.6 Data Analysis Methods

The general bioinformatical data analysis pipeline follows the same steps as described in {Eggers, 2021 #9375}.

2.6.1 Read Processing

For ChIP- and MNase-sequencing experiments the sequencing was performed with an Illumina HiSeq1500 by LAFUGA. Basecalling was performed by Illumina's RTA software, version 1.18.66.3. Demultiplexing was done by JE demultiplexer (Girardot et al., 2016) using the barcodes from the Illumina Index read files. Then the demultiplexed files were aligned to the *D. melanogaster* release 6 reference genome (BDGP6) using Bowtie2 (Langmead and Salzberg, 2012) version 2.2.9. (parameter "--end-to-end --very-sensitive --no-unal --no-mixed --no-discordant -I 10 -X 220") and filtered for quality using samtools 1.6 (Li et al., 2009) with a MAPQ score cutoff of -q 10, allowing only for high quality reads of sizes between 10 and 220 bp.

2.6.2 Replicate Correlation

To be able to summarize the replicates within the same experiment they had to be compared first. To this end, the reads for each replicate were formatted to ".bed" format using bedtools2 (Quinlan and Hall, 2010) by calling the function bamToBed and sampled to the same read count for normalization. Then the ".bed" files were imported to R and genome coverages were calculated. As mostly GA-binding proteins were to be compared, 5000 potential GA-rich binding sites, as determined by a FIMO search, were mapped in the genome to have a pool of bound and unbound sites to compare. Then I calculated the cumulative coverages of a 100 bp window around sites of interest for all ChIP-seq samples. Then these values were correlated by spearman correlation between the samples to affirm similarity between them. For MNase samples average dyad densities around sites of interest were plotted and compared for each replicate to affirm similarity. Here sites known to exhibit phased nucleosomal arrays served as the controls. If they were sufficiently similar (R>0.6 for ChIP and visual inspection for MNase samples) the sampled reads were summed up and used for further analysis. This allowed to avoid normalization against an input with possible zero values and to use the larger combined dataset for peak calling, thus improving the robustness of the resulting peaks.

2.6.3 Peak Calling

Peaks were called using Homer (Heinz et al., 2010) version 4.9.1 calling the functions makeTagDirectory (parameters -single -fragLength 150) and findPeaks (parameters -style factor -size 150 -F 8 -L 2).I used the corresponding negative samples, in which the IP was done without adding the respective protein as control. Only for NURF301, which is endogenously present in the DREX and could therefore not be excluded, the IPs were normalized against their inputs. Calling peaks against a negative IP control is preferable where possible, as this allows to account for antibody bias in immunoprecipitations. Peak calling was done with the accumulated reads from all replicates for each sample and their respective controls, resulting in more robust peaks through the additional coverage. The 309 HAS and 56 PionX regions were used for annotation and were chosen as defined by (Villa et al., 2016). CLAMP sites were defined by (Albig et al., 2019).

2.6.4 De novo Motif Discovery

To define the enriched motifs within each peak set the MEME suite was used (Bailey and Elkan, 1994) (version 4.11.4, parameters -mod anr -dna -revcomp -nmotifs 1). The MNase-assisted shearing leads to a situation similar to what is seen in ChIP-exo experiments, where the protein binding site is located directly at the beginning or end of a given read. Therefore, before analysis the peaks were resized to 200 bp to include 25 bp of sequences directly bordering the peaks to be sure to include potential flanking sequences.

2.6.5 Motif Search

Motif search using position weight matrixes from MEME in peak regions on the genome was performed with FIMO (Grant et al., 2011) version 5.0.2.

2.6.6 Browser Profiles

Browser profiles were generated using UCSCutils (http://genome.ucsc.edu.) version 3.4.1. calling the function makeUCSCfile using the summed up sample replicates Tag Directories, also used for the peak calling, and were normalized against the control. Values are fold change over control. Profiles were visualized using the IGV software (Robinson et al., 2011).

2.6.7 Data Analysis and Plotting

Data Analysis was conducted in R (R Core Team, 2014), using the tidyverse libraries (Wickham et al., 2019).

2.6.8 Venn Diagrams

Venn diagrams were made using the resized peaks as for the de novo motif discovery and allowing for a maximal gap between overlapping sites of 100 bp, effectively scoring sites as “overlapping” if their centers are separated by <1 peak width. Plots were drawn in R using the library Vennerable (https://github.com/js229/Vennerable).

2.6.9 Heatmaps and Cumulative Plots

Heatmaps were made using the R library “Complexheatmaps” (Gu et al., 2016) by selecting windows of 2000 bp around sites of interest of the calculated coverages normalized against a control, if applicable, and aligning them. The cumulative plots are made by calculating the mean of each column. Window identities are retained in the data and used for the annotation by overlapping them with the known HAS or the X chromosome.

2.6.10 Shape Analysis

To find and align all bound motifs within peaks, the peaks were extended by 25 bp on each side and the “Find Individual Motif Occurrences” (FIMO)(Grant et al., 2011) was employed (MEME suit version 5.0.2. and using parameters --qv-thresh --thresh 0.01). If multiple hits were recorded within one peak only the best hit was considered, to allow unbiased selection of weak motifs if no stronger motifs are nearby, while ignoring false positive overlapping motifs. False positives can happen as the motif is degenerate and repetitive allowing for multiple hits in a GA-rich region. Shapes were then calculated using DNAshapeR (Chiu et al., 2016), which is based on Monte-Carlo simulations to estimate approximate values for DNA Roll and Propeller Twist. Plots were drawn using ggplot2.

2.6.11 Data Repositories

The published data can be found at the Gene Expression Omnibus (GEO) under the tag GSE169222. Additionally, the R and Unix scripts developed for this thesis can be found on Github at https://github.com/nikolas848/eggers2021

3. Results

3.1 Characterization of Assembled Chromatin

3.1.1 Purification of Genomic DNA

Chromatin was reconstituted in vitro using plasmids or other small nucleotide sequences before, demonstrating that using DREX extracts also whole genomes can be chromatinized in vitro (Baldi et al., 2018b). This study aimed to enhance the scope of previous reconstitutions and also aimed to provide a physiological substrate for transcription factor binding. Using the whole genome enables an unbiased analysis and introduces unspecific binding competition. Genomic DNA (gDNA) was purified from BG-3 cells, because these are diploid male cells which eases computational analysis. After isolation of the DNA using the Qiagen kits (2.3.2), the DNA was probed by gel electrophoresis. The gDNA showed a high molecular weight band above 10 kbp and more DNA was detected within the gel pocket hinting to even larger fragments (Figure 14). The manufacturer of the kit used assumes fragments to be between 100-150 kbp. There is also a weak and broad band of smaller (<0.5 kb) molecular weight fragments which could stem from RNA in the sample or from DNA degradation.

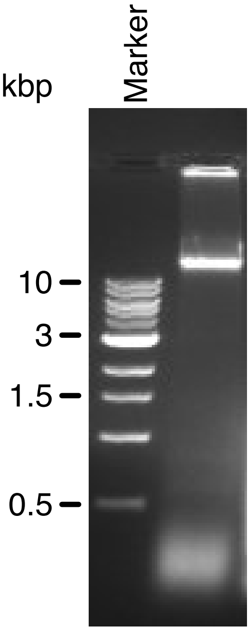


Figure 14: Purified genomic DNA fragments using standard kits are high molecular weight. DNA was purified and separated through gel electrophoresis. DNA marker sizes in kbp are annotated.

3.1.2 Assembly / Extract Titration

This study was conceived to test transcription factor targeting in vitro in an as physiological setting as possible. Chromatin was reconstituted using Drosophila embryonal extracts to provide the tested transcription factors with a physiological binding substrate. The protocol and the factors as described in 2.5.3 were used to assemble chromatin.

To test how successful the assemblies were, the chromatin was partially digested with micrococcal nuclease (2.5.4) and the purified DNA was assayed by gel electrophoresis. If nucleosomes are loaded efficiently onto the DNA this results in the formation of characteristic DNA “ladders”, as each nucleosome protects about 147 bp of DNA from digestion. Multiples of this value representing di-, tri-, tetra- and oligo-nucleosomes are present if nucleosomes are regularly spaced. The ladder serves as a quality control of assembly degree. MNase digestion ladders were titrated using 1 µg of DNA and increasing amounts of DREX for each DREX and DNA purification to assess the necessary amounts of each for proper assembly. Here the extend of the DNA ladder increases with increasing DREX amounts, suggesting a saturation of this particular DNA sample at 80 µl of extract per µg of genomic DNA (Figure 15).

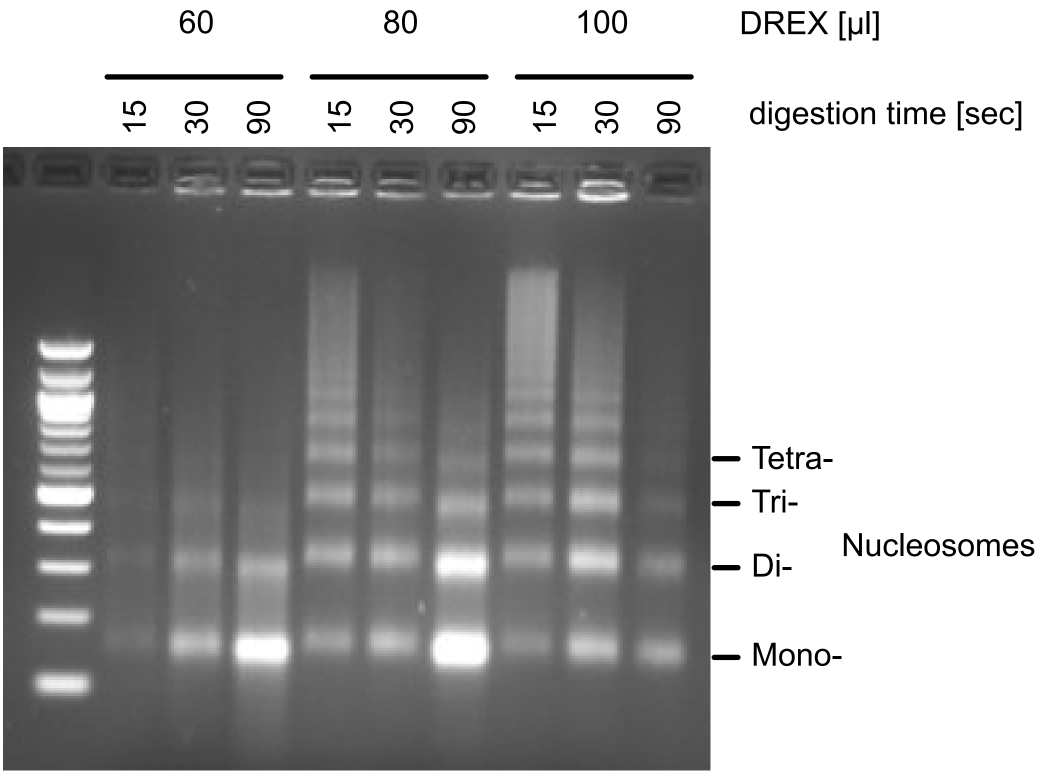


Figure 15: MNase-digested DREX-assembled chromatin forms fragment ladders. Chromatin was assembled in the presence of the indicated amounts of DREX and digested with MNase for the indicated times. Protein concentrations of DREX extracts are usually ~40 mg/ml. The resulting fragments were visualized after agarose gel electrophoresis by ethidium bromide staining. The migration of oligo-nucleosome-sized fragments is indicated.

3.1.3 Mapping Nucleosomes

To assess the positions of nucleosomes genome-wide it is necessary to map the nucleosome dyads. To this end, the mononucleosomal-sized fragments from gel-electrophoretically separated MNase-digested chromatin samples were excised from the gel and purified. After paired-end sequencing, the sequence tags are aligned to the reference genome and used to determine the central base pair for each read pair. For nucleosomal maps only fragments with a size of 125 to 220 bp were selected for further analysis, this excludes non-mononucleosomal fragments. The remaining midpoints for each read pair represent the nucleosome dyad position. Using these uniformly weighted full reads often results in spurious occupancy maps due to overlapping nucleosomes from the alternative positions found in the mixed population of the sample used. Therefore, often a center-weighted method is used where each nucleosome dyad is mapped and the read is then resized. A value of 50 bp per read was used to improve visualization of the coverage in the analyzes. All these resized reads are summarized and normalized to the number of reads present in each sample. This normalized read count average per base pair is termed coverage and represents the relative nucleosome occupancy at a given genomic location.

3.1.4 Phasing at Boundary Factor Binding Sites

Using this normalized coverage, the nucleosomes can be visualized especially well at genomic locations with highly positioned nucleosomes. In DREX-reconstituted chromatin certain boundary factors bind the DNA tightly, which leads to nucleosome remodeler-mediated phasing of nucleosomes against this boundary. At binding sites for Phaser and Su(Hw) this leads to phased nucleosomal arrays which can be visualized individually as a heatmap or by summarizing the mean of all known binding sites. Visualizing all sites individually in a heatmap is advantageous as the contribution of each site to the overall mean can be determined. The number of sites with uniform occupancy is difficult to determine from the means alone.

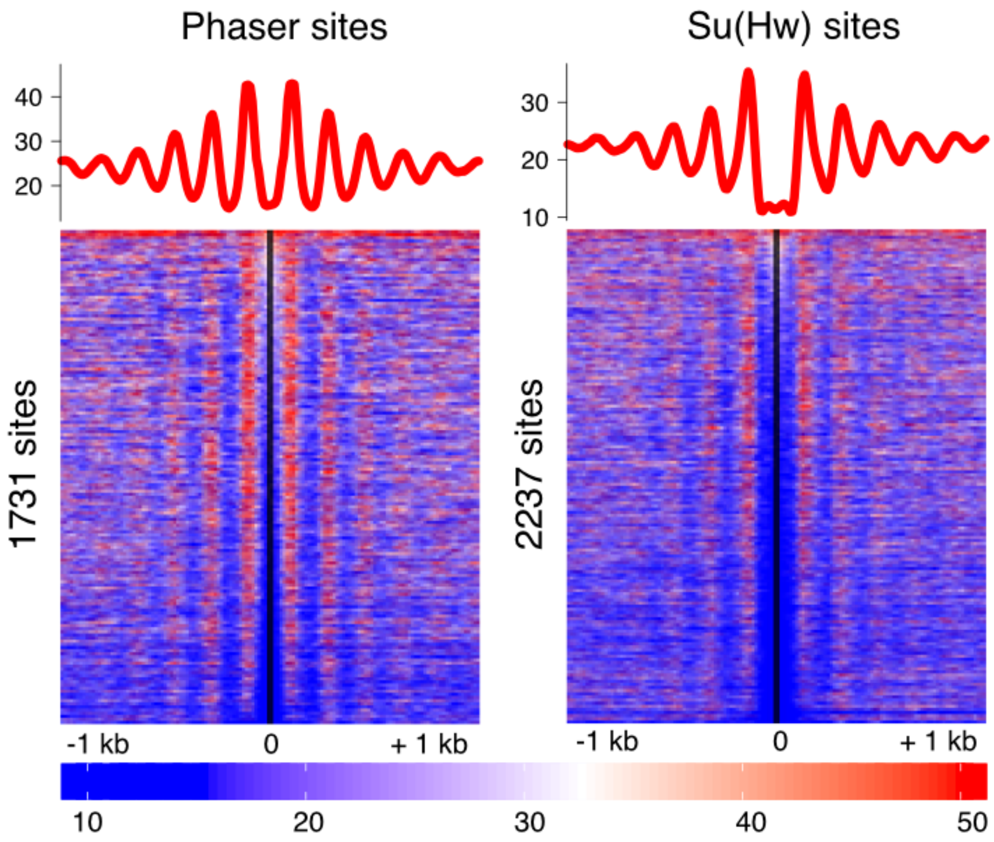


Figure 16: Nucleosomes form phased nucleosomal arrays around boundary sites. Chromatin was assembled and nucleosomes were mapped using MNase-seq. At known Phaser and and Su(Hw) binding sites and coverage windows of 2000 bp around the sites were cut out, aligned and the mean for each column calculated. Data was illustrated by average profiles (top) and heatmaps of individual regions (bottom). The profiles summarize the results of 2 biological replicates (n=2).

3.1.5 H1 Increases Linker Length

DREX-assembled chromatin mostly lacks the canonical linker histone H1 and instead contains mostly the linker binding proteins HMG-D and bigH1 (Climent-Canto et al., 2020; Ner and Travers, 1994). These proteins bind nucleosomes in a similar fashion but such chromatin might have different properties and may be particularly open for the transcription factors that orchestrate the first wave of zygotic transcription. Earlier experiments demonstrated that purified H1 can be faithfully incorporated into plasmid chromatin and replace BigH1, if added to DREX during the assembly (Becker and Wu, 1992; Ner et al., 2001; Sandaltzopoulos et al., 1994).

To test if H1 incorporates into chromatin in vitro at a genome-wide level as expected, H1 was titrated into the assemblies. Using physiological levels of H1 leads to the expected increase in nucleosome repeat length (NRL) in bulk chromatin (Figure 17)(Blank and Becker, 1995; Sandaltzopoulos et al., 1994) of about 25 bp. This was true up to a point where high H1 concentrations seem to interfere with the protocol leading to smeared bands. This may be due to non-stochiometric and non-physiological association of H1.

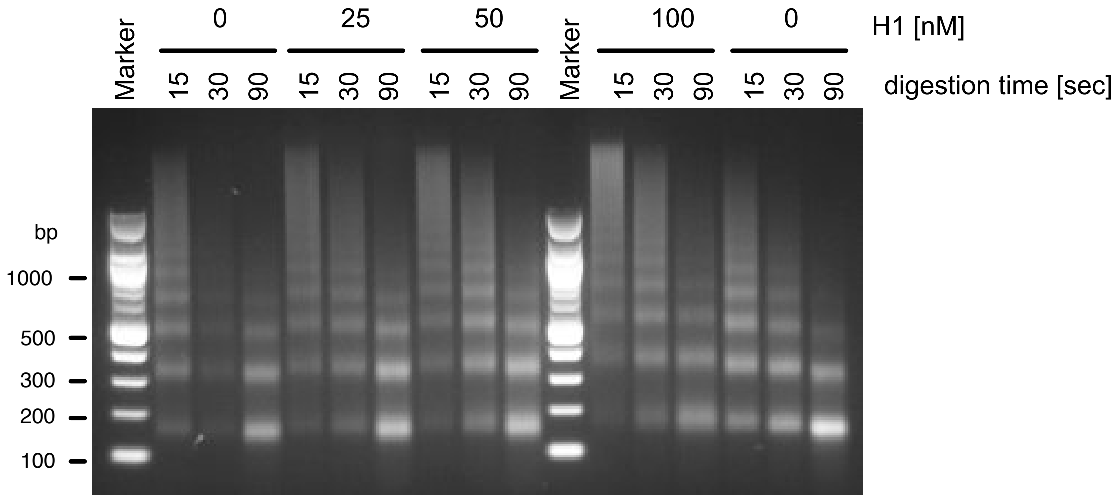


Figure 17: H1 incorporation increases the linker length. Chromatin was assembled in the presence of the indicated amounts of histone H1 and then digested with MNase for the indicated times. The resulting fragments were visualized after agarose gel electrophoresis by ethidium bromide staining. Marker: DNA fragment sizes in bp.

Looking at electrophoretically separated bands can only visualize bulk chromatin and nucleosome sizes. To appreciate the difference H1 makes on linker lengths more specifically the coverages from MNase sequencing can be aligned to particular points of interest, such as Phaser binding sites. The coverage of 1731 such sites was merged and the mean and standard deviation plotted for each sample with different H1 concentrations.

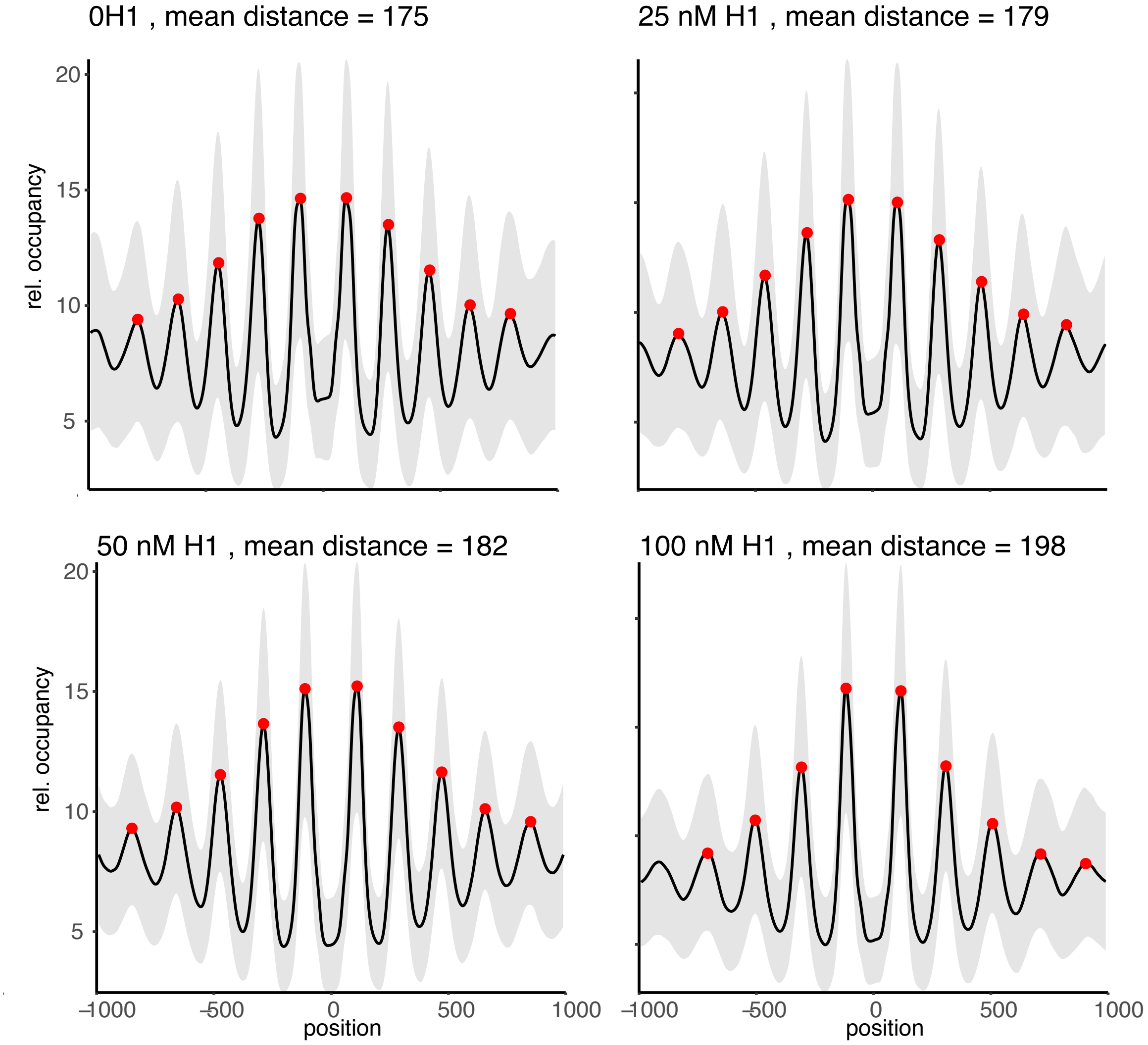


Figure 18: H1 shifts nucleosome positions around boundary factors. Chromatin was reconstituted with the indicated amounts of H1 and nucleosomes were mapped by MNase-seq. Coverage windows of 2000 bp around 1731 known Phaser binding sites were cut out, aligned and summarized. Mean and SD are plotted and dyads annotated in red.

After plotting of the curve, the maxima were determined and the distances between them were calculated. The protein footprint in the middle was excluded from the analysis as it differs in size. The data shows an increase of mean dyad distance from 175 bp for the control to 198 bp at 100 nM H1 with intermediate values for the intermediate H1 concentrations. Subtracting 147 bp of directly bound nucleosomal DNA this leaves a linker length of 28 bp for the control and 49 bp in the presence of 100 nM H1 (Figure 18).

This increase is in line with values from the literature and demonstrates the physiological way in which DREX-derived chromatin adapts to changes in protein composition during assembly. This analysis allowed, for the first time, to compare nucleosome positions genome-wide in the absence and presence of the linker histone. This would have been difficult in vivo settings due to the essential nature of H1.

3.1.6 Damage Response

Previous studies in our lab established that DREX-assembled chromatin has in vivo like ability to sense and respond to double-strand DNA breaks (Harpprecht et al., 2019). This includes the incorporation of the histone variant H2A.V into chromatin and more importantly the C-terminal phosphorylation of H2A.V as this modification is essential for the DNA damage response in Drosophila, similar to the C-terminal H2A.X phosphorylation in mammals (Harpprecht et al., 2019; Madigan et al., 2002).

The gDNA used in the DREX-assisted chromatin assemblies is randomly broken into ~100 kb large fragments. To evaluate if this has an effect on chromatin multiple known double-strand break locations would have to be summarized. To this end, defined breaks were introduced by restriction enzymes, so that the signal from these sites could be summed up to maximize the potential signal. It was necessary to find a balance between as few cut sites as possible not to dilute the damage response signal and enough cut sites to acquire sufficient signal to distinguish it from the control. Therefore, DNA restriction enzymes were queried for their number of cut sites within the Drosophila genome. I aimed to find a restriction enzyme that has at least 1000 cut sites to allow statistical analysis, resulting in chromatin fragments larger than 100 kb (Figure 19).

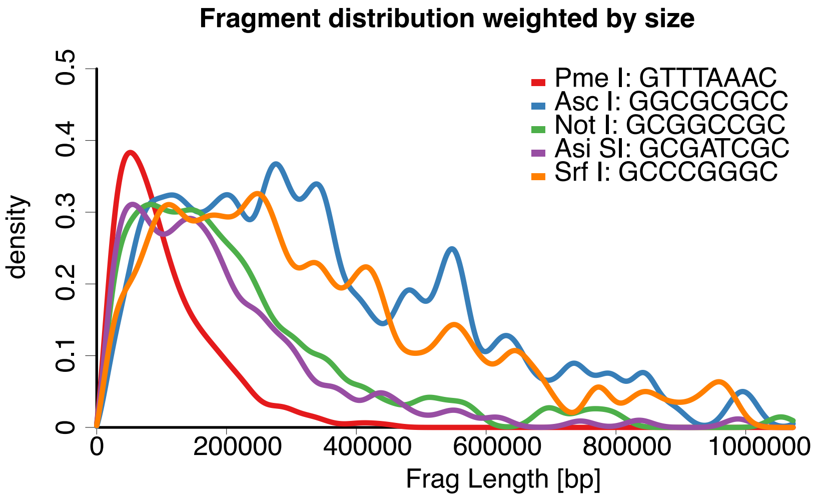


Figure 19: Fragment distribution of selected restriction enzymes. Fragments were calculated bioinformatically and the density was scaled to fragment size at each fragment length position.

The computational analysis suggests that NotI and AscI digestion would lead to the desired fragment sizes. To obtain a mostly unbroken DNA sample, genomic DNA was purified from BG-3 cells by first lysing the cells with SDS and Proteinase K. Then phenol:chloroform:isoamyl-alcohol was added followed by centrifugation and ethanol precipitation. The purified DNA was assembled into chromatin and digested with the candidate proteins. The overall size composition of genomic DNA was not altered by the restriction with these rare cutters (Figure 20, left). To probe how efficient restriction was a region of interest was measured by qPCR using primers that span a single known cut site. Using this approach, the data shows that AscI did not perform as expected, but NotI had high restriction efficiency at its respective cut site (Figure 20, right).

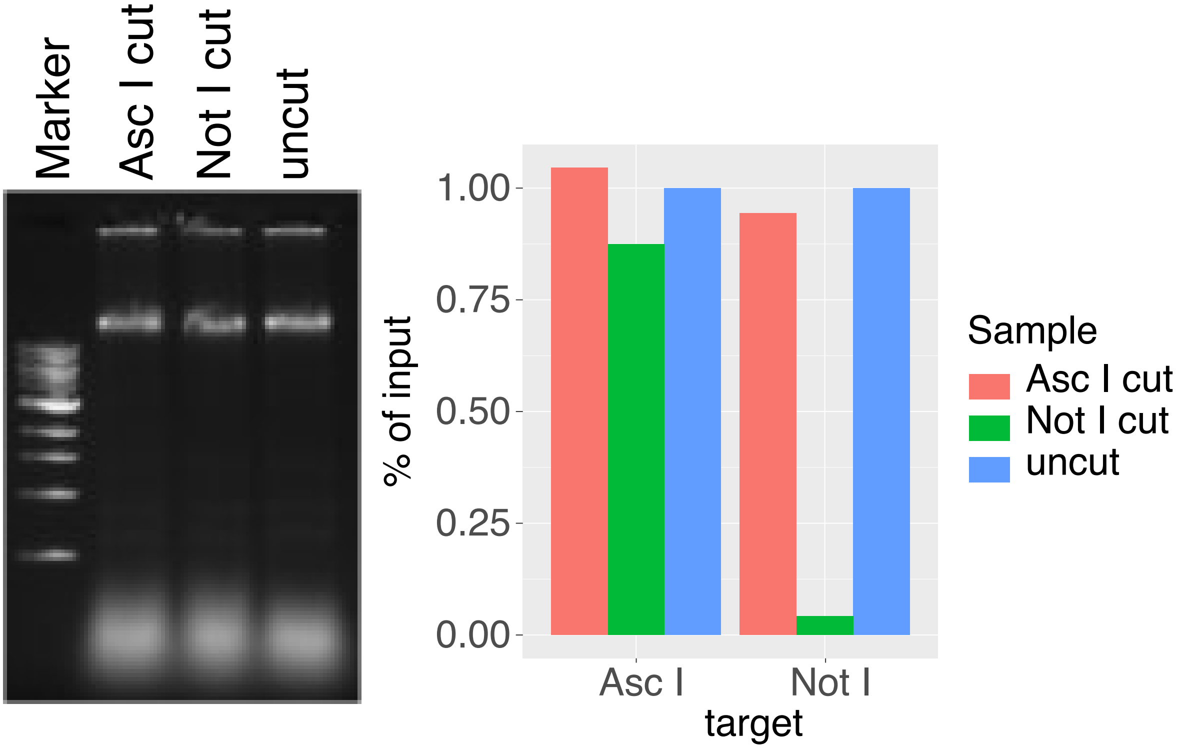


Figure 20: Not I cuts gDNA efficiently without changing overall composition. Left: gDNA was digested with the annotated restriction enzyme and resulting fragments were visualized after agarose gel electrophoresis by ethidium bromide staining. Right: qPCR analysis of ASCI and NotI restriction efficiency using digested fragments for a PCR with Primers spanning a known cut site.

Using chromatin bearing NotI cuts, I mapped H2A.V and H3 by MNase-seq. Mapping the coverage at known Not I binding sites revealed that most of the sites are efficiently cleaved. This approach also reveals a considerable resection of the DNA at the cut sites (Figure 21). The removal of nucleosomes and resection of DNA at double-strand breaks was suggested earlier and seems to be an intrinsic property of the DREX extract (Harpprecht et al., 2019). H2A.V was not enriched over the H3 at DSB. As the site-directed damage might influence downstream analysis it seemed prudent to use the undigested gDNA bearing random DNA breaks, as any resection effects would be averaged out within the heterogeneous population.

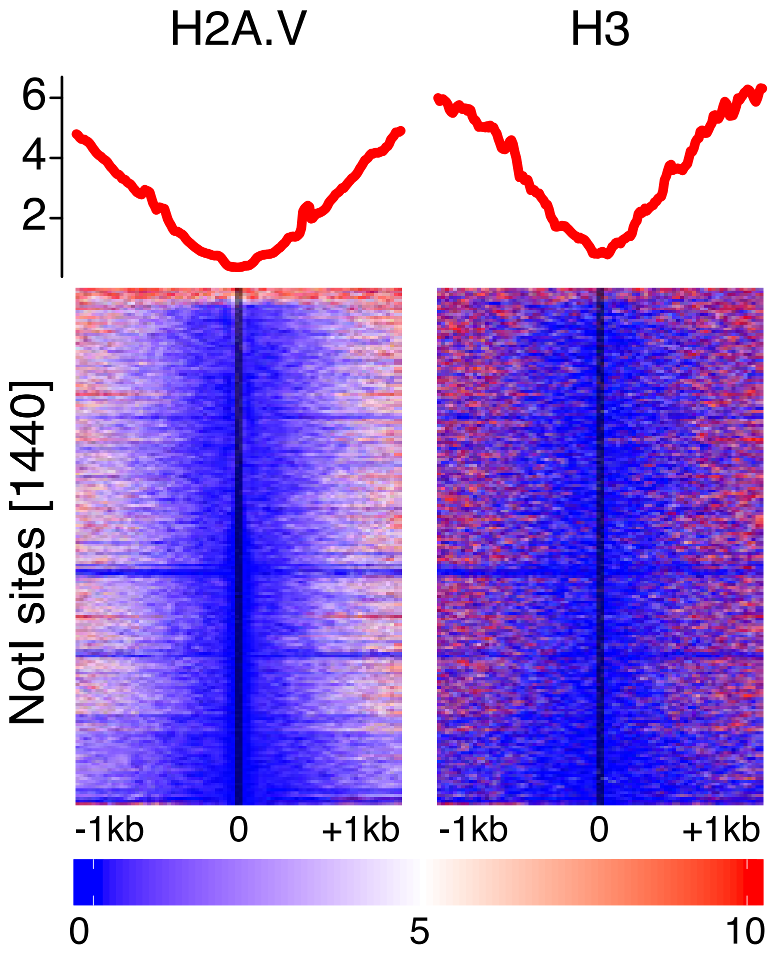


Figure 21: DNA is resected around double-strand breaks. gDNA was digested by NotI and then used for chromatin assembly. Nucleosome positions were determined by ChIP-seq for the annotated histones. Coverage windows of 2000 bp were selected around 1440 known NotI cut sites and aligned. Individual sites were plotted as a heatmap and the mean was plotted above.

3.1.7 PPPS

Recent studies published during the time frame of this work in the field of nucleosomal organization suggested that cells organize their inner compartments through liquid-liquid phase separation (LLPS) or, in the case of long polymers such as chromatin, polymer-polymer phase separation (PPPS). This leads to chemically distinct so-called condensates. Chromatin was also implicated to instrumentalize this transition to aid 3-dimensional organization of chromosomes (Zenk et al., 2021). To test wether in vitro DREX-reconstituted chromatin can recapitulate the findings made in vivo, reconstitutes were imaged by fluorescent microscopy. Chromatin was assembled for 4 h as described in 2.5.3 and then stained with Sir-DNA, a DNA-intercalating staining agent, which is orders of magnitude more fluorescent if bound to DNA. This made washing away of unbound agent unnecessary, which would have been difficult as chromatin is not as easily fixed to a microscopy slide as cells. Images were taken as outlined in 2.5.9.

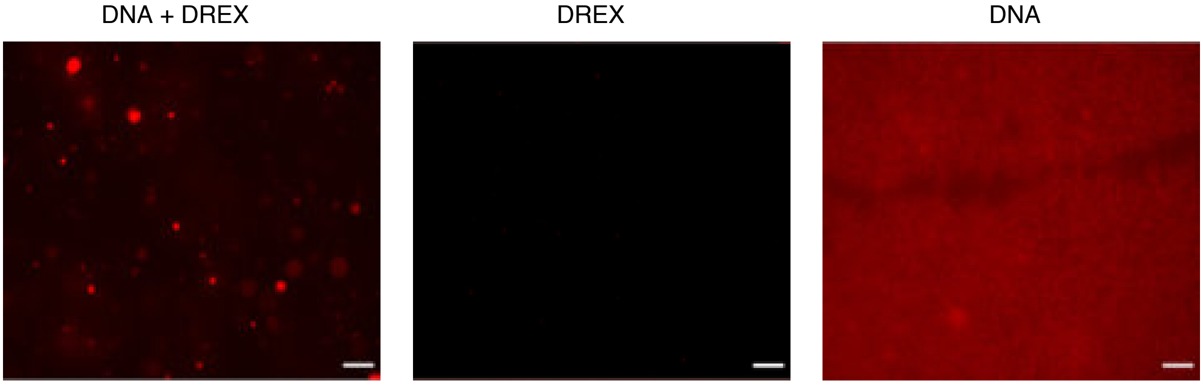


Figure 22: Chromatin assembled by DREX forms condensates in vitro. Chromatin was assembled as described and incubated for 4 hours, either with DREX and DNA (left), omitting DNA (center panel), or omitting DREX (right panel). Then samples were stained with the SiR-DNA Kit for 15 min at RT. Fluorescence was imaged on a confocal microscope at the Cy5 channel as described in 2.5.9. Scale bar represents 10 µM. Adapted and reprinted with permission from Oxford University Press (Eggers and Becker, 2021).

The assembly reaction omitting DREX (labeled DNA), shows broadly distributed signal as expected, while the sample lacking DNA (DREX) gives little signal as no DNA is present (Figure 22). Properly assembled chromatin on the other hand forms globules of 1-10 µm in diameter, approximating a nucleus in size, as expected from polymer theory and previous reports (Erdel and Rippe, 2018; Gibson et al., 2019a; Maeshima et al., 2016) As the formation of chromatin globules in vivo and vitro underly the same physical rules and are both dominated by the nucleosome-nucleosome interactions in cis (Chodaparambil et al., 2007), the local DNA concentration in these condensates can be assumed to be similar to preblastoderm nuclei. This further strengthens the assumption that DREX-reconstituted chromatin is an appropriate substrate to test transcription factor-chromatin interactions. To our knowledge, this is the first study demonstrating PPPS using a full metazoan genome in vitro.

3.2 Optimization of IP Protocols

3.2.1 Protein Preparations

The goal of this work was to reconstitute chromatin assembly and transcription factor binding in a cell free system. To be able to manipulate each relevant factor, first all pieces of the puzzle needed to be purified individually. The quality and purity of factor preparations are of crucial importance. Therefore, the purifications were included in the results to highlight the quality of the components used in the later analyses.

Proteins were purified from cell cultures infected with baculovirus expression vectors according to the respective protocols in the methods 2.4.2 - 2.4.5. After purification, protein preparations were subjected to gel electrophoresis and stained by Coomassie-Blue.

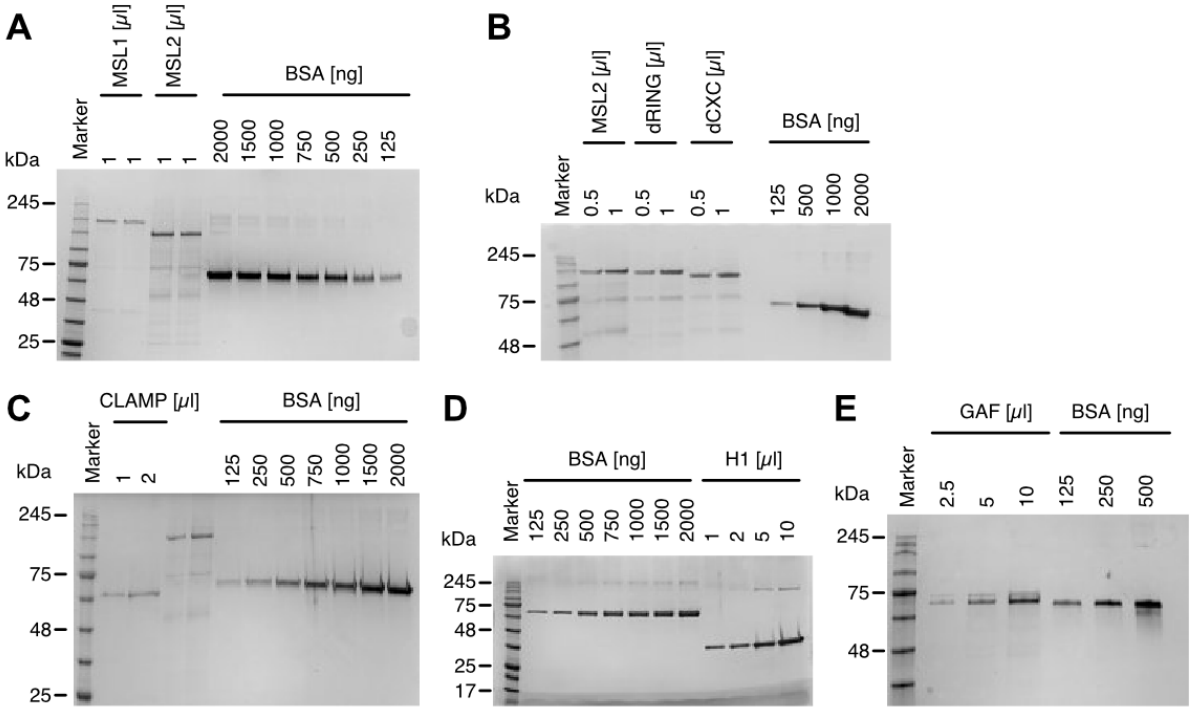


Figure 23: SDS-PAGE analysis (Coomassie brilliant blue staining) of protein purifications. Proteins were purified according to protocol, separated by gel electrophoresis on a 10% SDS-gel and stained by Coomassie Blue. Markers and sizes are annotated on the left of each gel. Protein samples and amounts added [in µl] are annotated above and concentrations were estimated using a BSA standard as comparison. Proteins measured were in A: MSL1 and MSL2; B: MSL2 mutants; C: CLAMP, D: H1; and in E: GAF.

To approximate the protein concentrations a titration of BSA standards was used. Most of the proteins were purified by FLAG-affinity purification. The masses of the FLAG-tagged proteins were: for MSL1 118 kDa, MSL2 84 kDa, CLAMP 62 kDa, GAF 64 kDa. H1 on the other hand was purified through phenyl sepharose column purification. Its size was roughly 26 kDa. It is noteworthy that MSL proteins run at an apparently higher molecular weight than would be expected from their size. Lanes containing MSL2 always showed some weaker bands of lower molecular weight, possibly degradation products (Figure 23, A). The MSL2 mutants run at slightly lower molecular weights than the wildtype, as expected.

Increasing the concentration of MSL2, GAF or CLAMP further using e.g. Amicon tubes, led to a massive loss in overall protein amount, while not increasing the protein concentrations much. These proteins seem to be unsuitable for this kind of protocol and this procedure is therefore not recommended. Proteins were used “as is” at the concentrations determined by the SDS-PAGE analysis.

3.2.2 Western blots of relevant factors

DREX is prepared from preblastoderm extracts lacking transcription. However, some proteins, such as CLAMP and GAF, are maternally deposited as proteins and mRNA. Both these factors are necessary for ZGA and therefore present and active, however in relatively small amounts. The DREX purification protocol separates the nuclear fraction, including the chromatin and factors bound to it, from the later used extract, and only minimal concentrations of these factors are present before the onset of ZGA, so it was possible that the extract would be functionally free of them. It was, therefore, necessary to monitor the endogenous levels of these proteins.

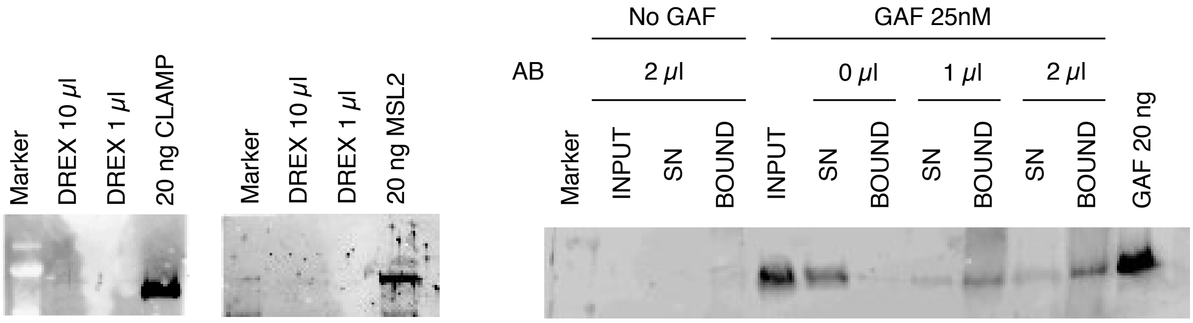


Figure 24: Western blots of relevant factors showing that DREX contains no detectable CLAMP, MSL2 or GAF. Western blots probing for each factor in the indicated amounts of DREX. Addition of annotated amount of recombinant protein serves as control. For GAF also the IP efficiency of the used antibody is shown, while for the other two publications demonstrating this are available (Albig et al., 2019). Adapted and reprinted with permission from Oxford University Press (Eggers and Becker, 2021).

To this end, functional assays can be performed testing nucleosomal distributions around known binding sites and these are performed as controls for the later ChIP-sequencing results. However, also Western blots were made using specific antibodies for the CLAMP, GAF, and MSL2 to measure their levels within DREX (Figure 24). None of the factors were detected in any significant concentrations within DREX. This allowed us to titrate them freely in to the extracts and using IPs without added protein as negative controls in the bioinformatical analysis of the ChIP-seqs. This is an advantage over simply using an input, as usually necessary in vivo experiments as producing a null mutant is laborious or impossible, because it accounts for possible antibody bias within the sample.

3.2.3 EDTA Reduces MSL2 Binding

Many transcription factors have a zinc finger DNA binding domain, in which Zn^{2+} ions are central to a structural domain that mediates the DNA interaction. Indeed, also MSL2, CLAMP and GAF utilize zinc finger domains for DNA binding. All DNA binding reactions were therefore supplemented with 10 µM $ZnCl_2$ to allow proper folding and functionality of these proteins. Nevertheless, early experiments yielded only low IP efficiencies and troubleshooting illustrated a known but underappreciated property of EDTA (and EGTA). The zinc-binding motifs found in transcription factors were found to have dissociation constants (K_D) in the range of 10^{-9}–10^{-11} M, while EDTA was shown to bind zinc with a dissociation constants of approximately 10^{-16} M (Michael et al., 1992). Therefore, EDTA will easily outcompete the proteins and deplete the solution of free Zn^{2+} and might interfere with proper protein folding and function (Nyborg and Peersen, 2004). To demonstrate the effects of EDTA on IP efficiency the standard MSL2 DNA immunoprecipitation assay was employed in presence and absence of EDTA using 50 or 100 nM of MSL2. The resulting DNA fragments were measured by qPCR (Figure 25).

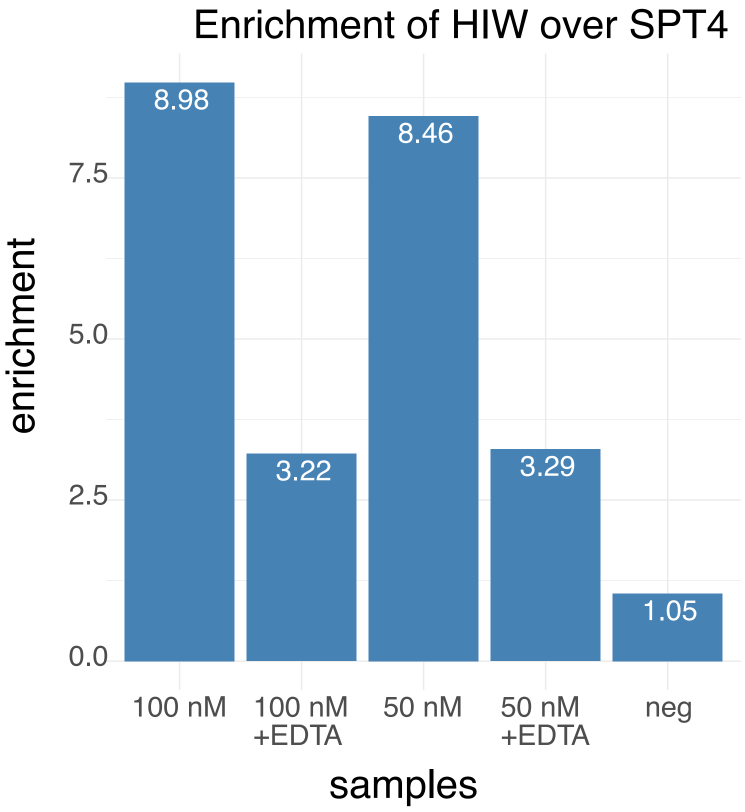


Figure 25: EDTA reduces enrichment of HAS as determined by DIP-qPCR of MSL2. Enrichment was determined by delta-delta-CT of the MSL2 IP of a HAS region (HIW) against the input and a control region (SPT4). Concentrations of MSL2 are annotated and the presence of 0.5 mM EDTA is indicated.

The enrichment of target sites was much lower in presence of 0.5 mM EDTA. Consequently, EDTA was excluded from all buffers contributing to the chromatin assemblies and ChIP reactions before the crosslinking step to allow the transcription factors to bind to chromatin without interference. It is advisable to keep this known, but underappreciated, property of chelating factors in mind when working with transcription factors in vitro.

3.3 ChIP experiments

3.3.1 Chip Protocol

This study is the first to test transcription factor binding on in vitro-reconstituted chromatin on a genome-wide scale. As a proof of principle, I first established the binding patterns of MSL2 and CLAMP to reconstituted chromatin, two well characterized DNA binding proteins previously tested for DNA binding in vivo and in DIP assays (Albig et al., 2019; Villa et al., 2016). This provides reliable reference profiles to compare the chromatin binding to. As nucleosomes occlude most of the DNA it is not obvious that these proteins would bind in the same fashion to naked DNA as to chromatin. For example, only 30% of CLAMP binding sites as discovered by DIP overlap with the in vivo peaks from previous studies. It was therefore hypothesized that much of this false-positive binding could be inhibited by nucleosome competition in vitro (Figure 28, left). Similarly, for MSL2, the DIP and in vivo profiles (Figure 28, right) are starkly different. MSL2 binds about 250 sites in vitro but only ~50% of these binding events are located on the X chromosome (Figure 29). To test if the occlusion of false positive sites by nucleosome competition would be enough to change the binding profiles to a more physiological pattern, I employed a novel in vitro ChIP protocol as depicted in Figure 26.

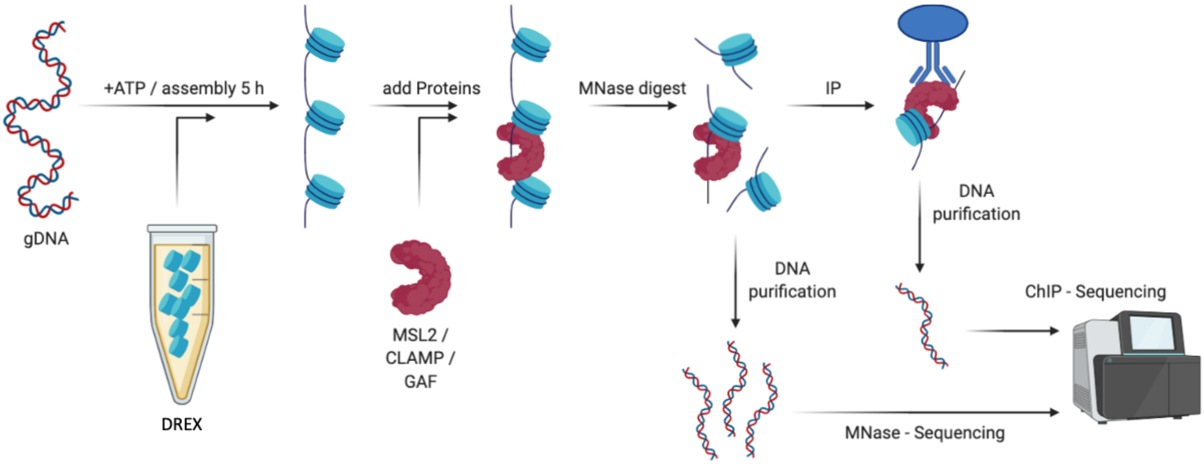


Figure 26: Schematic depiction of the employed cell-free chromatin immunoprecipitation protocol. In short, genomic DNA is assembled into chromatin for 4 h using DREX in presence of ATP, then proteins of interest are added, allowed to bind for 1 h and fixed by 5% formaldehyde. Samples are digested by MNase and either subjected to MNase-seq or IP followed by ChIP-seq. Created with BioRender.com.

Chromatin was assembled using DREX for 4 hours then proteins were added and allowed to bind for another hour before crosslinking with 3.7% formaldehyde for 5 minutes. The chromatin was digested with MNase, then CLAMP and MSL2 binding were mapped by ChIP-seq using specific antibodies.

To assure high quality of the datasets, multiple replicates were used for each experiment. For the in vitro setting it was determined that a replicate would involve an independent protein purification, DREX extract and genomic DNA purification. Additionally, replicates were done on different days.

3.3.2 Peak Calling

Peak calling is a bioinformatic method to determine areas in the genome that are enriched in reads received from a ChIP-seq experiment. It is used to determine the genomic locations of DNA-interacting molecules. Peal calling employs software programs such as Homer or MACS, which are specialized on the signals received from ChIP-seq (Heinz et al., 2010; Zhang et al., 2008). A potential "peak" must be enriched over the general background of the control and additionally above the local background of the sample itself. There are no precise rules for how much a signal must be enriched to be considered a peak and precise settings for each experimental design and DNA interacting molecule have to be tuned to faithfully "call" genuinely enriched peaks. The biggest challenges for peak calling are low specific signal intensities over a high background signal (for ChIP-seq samples usually <2% of reads are located within the peaks) and possible technical artifacts, which lead to enrichment of certain sequences in sequencing libraries. Still peak calling is an inherently binary system, which often fails to reflect the subtleties found in biology.

It is problematic to call peaks in areas where the background is low as the negative control or input over which the sample peak might be enriched, might lack the coverage necessary for faithful calculations. The algorithm cannot give enrichment values for genomic locations where it would need to divide by 0. It is therefore desirable to use as many reads as possible for each sample and control as this strengthens the data on which conclusions are made.

Another complication is introduced by the integration of multiple replicates. One way to sum up replicates is to call peaks for each replicate individually and then only consider the ones present in all or most of them. This can be problematic for signals that are present (by visual inspection of browser profiles), but below the set threshold to be considered a peak in some of the replicates, as they will be discarded together with peaks of low reliability that are found in just a single replicate. Also, the number of reads used for peak calling is necessarily only a fraction of those of all replicates.

Another possibility to integrate replicates is to add up their reads, but as each replicate has a different number of reads for technical reasons, this necessitates normalization by the number of reads found in each sample, to avoid putting excessive weight to the replicate with the most reads. This leads to fractionated reads which introduce new problems for the peak calling algorithm.

To avoid normalization to the read count and additionally the input, which can be problematic around sites with no reads in the input, I decided to randomly sample the reads from each replicate to match the number of reads found in the replicate with the lowest read number and to add them up. The same was done for the inputs or negative controls. This allows to sum up replicates without normalization and without giving excessive weight to the replicates with the most reads. Then peak calling was performed on this larger dataset. This is advantageous as I can draw from a larger number of reads for peak calling, leaving fewer gaps in the genomic coverage, while at the same time giving more weight to peaks found in all replicates.

On the other hand, this introduces new bias as a very strong peak found in only one replicate will be called. This replicate-mean-centered workflow cannot easily distinguish the contributions made by each of the replicates. Also, by subsampling, some reads are lost from the dataset. This is only problematic if the samples have massively differing read counts. If not, the summation rather strengthens the dataset by broadening the number of reads used for downstream analysis.

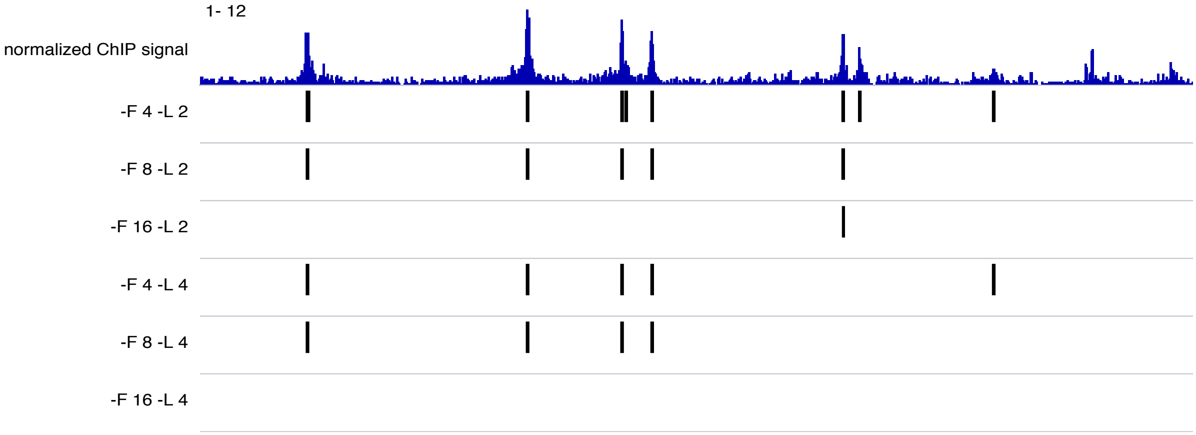


Figure 27: Peak calling at a genomic locus with a complex coverage profile, showcasing only some options to “call peaks”. ChIP calling options are annotated: F represents the necessary “factor” enrichment over input, while L represents the necessary enrichment over the “local” background to be considered a peak. All peaks were called as 150 bp windows and with one peak width minimal distance between peaks.

The general settings I chose for peak calling were: a sequence window of 150 bp per peak, with 8-fold enrichment of the sample reads over the control and 2-fold enrichment over the local background of the IP experiment with a minimal distance of one peak size between two peaks. This resulted in the most faithful representation of bioinformatically called peaks to the genomic coverage as seen by visual inspection. Still, one should keep in mind that peaks are a necessarily binary information which might hide the subtleties found in the raw data. For most analyses however it is not necessary to actually call every peak in a dataset. Rather, a subset of representative peaks can serve as an approximation to the majority of binding events.

3.3.3 Estimation of Physiological Concentrations

To estimate the necessary concentrations for the in vitro experiments the following was assumed: Using a mean weight of 660 g/mol per base pair for a ~160 million base pairs male fly genome, a typical ChIP-seq experiment using 1 µg of gDNA has about ~6*10^6 copies of that genome. Assuming ~500 binding sites per genome for MSL2, there are ~3*10^9 binding sites in each male genome and twice as many in a female genome. As the used gDNA is a from a male cell line a total of ~3*10^9 binding sites can be assumed. In male cells in vivo, a stochiometric ratio of about 2:1 to 3:1 MSL2 molecules per genomic binding site was measured (Bonnet et al., 2019). Ideally, the in vitro reconstitution would be performed with a concentration of ~33 pM MSL2 in 150 µl of IP sample using 1 µg of gDNA to achieve physiological ratios. This concentration, however, is not feasible for technical reasons. It was necessary to use a higher concentration of MSL2 to reliably produce functional sequencing libraries, while keeping the concentration to a minimum, to approximate the physiological levels and to not over saturate the binding sites. 25 nM was the lowest concentration that reliably yielded high-quality sequencing libraries for MSL2 and was therefore used as a reference in most IP experiments to ease comparability.

For a typical ChIP-seq sample of 150 µl using 25 nM of protein this translates to about $2.2*10^{12}$ molecules of that protein. Therefore, in the case of MSL2 there are about 750 molecules of MSL2 present per binding site using 1 µg DNA at a concentration of 25 nM MSL2. However, each of the binding sites can harbor multiple MREs, a FIMO search suggests ~3 per site, and MSL2 is assumed to bind as a dimer to the DNA (Hallacli et al., 2012), so the actual ration of MSL2 dimers per MRE could be closer to 125:1.

As seen in 3.1.7, the DNA and consequently the binding sites are however not equally distributed throughout the IP sample. If the proteins are, the actual ratio of proteins per binding site at the binding site locations would again be considerably closer to the physiological levels than the raw values suggest.

3.3.4 MSL2 and CLAMP Bind Chromatin in vitro

In piloting experiments it was established which amount of MSL2 and CLAMP was necessary for optimal in vitro sequencing results. This was necessary as low protein concentration result in bad sequencing data quality, while too high concentrations might result in binding site saturation and resulting nonspecific binding (Demakova et al., 2003; Villa et al., 2012). Aiming to add an as low as possible concentration of recombinant MSL2 and CLAMP to not oversaturate the sites, 25 nM was established as the best concentration to guarantee high quality ChIP-seq datasets and used in the following experiments. This concentration still represents a higher than physiological ratio of MSL2 molecules over HAS (see 3.3.3). Looking at the Venn diagrams of peak overlaps it can be seen that CLAMP binds to chromatin quite similarly as to DNA. Almost all ChIP-seq peaks overlap with the DIP-seq peaks (1004 of 1090, 93%, Figure 28), so the chromatinization only reduced the number of total binding sites to about 30%, while not influencing the general peak distribution. Still, some binding at most of the known DIP sites remains (see Figure 32, left). Another way of representing the data is to look at the chromosomal distribution of the peaks, this is especially relevant for MSL2 as all its functional binding sites reside on the X chromosome by definition. X-chromosomal enrichment of CLAMP peaks was slightly above the random distribution represented by the genome size and also slightly above the occurrence of GA-rich sites, especially in ChIP. This was different for MSL2. After chromatinization of the substrate the data reveals a similar degree of reduction in overall binding from 380 to 131 sites, similar to what was shown for CLAMP. But here there was a shift in peak distribution onto the X chromosome improving the X chromosomal enrichment from 45% to about 75%. For MSL2, it seems chromatinization and the accompanying nucleosome competition leads to preferred binding to physiologically relevant sites.

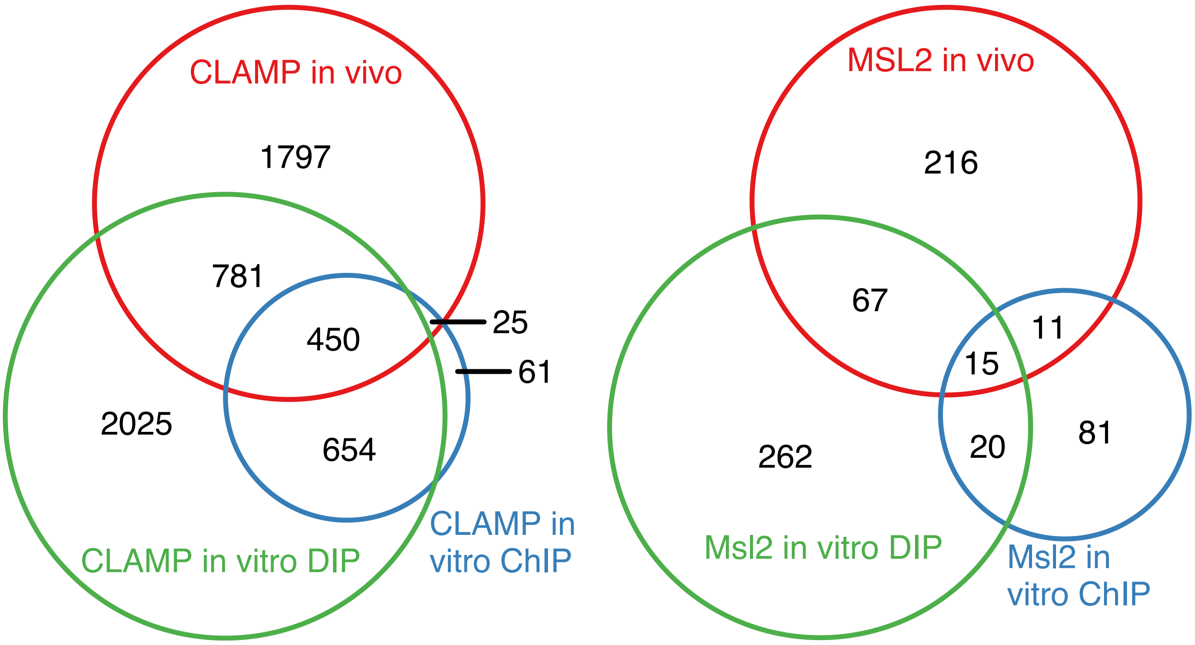


Figure 28: Venn diagrams of peak overlaps under different conditions. Adapted and reprinted with permission from Oxford University Press (Eggers and Becker, 2021).

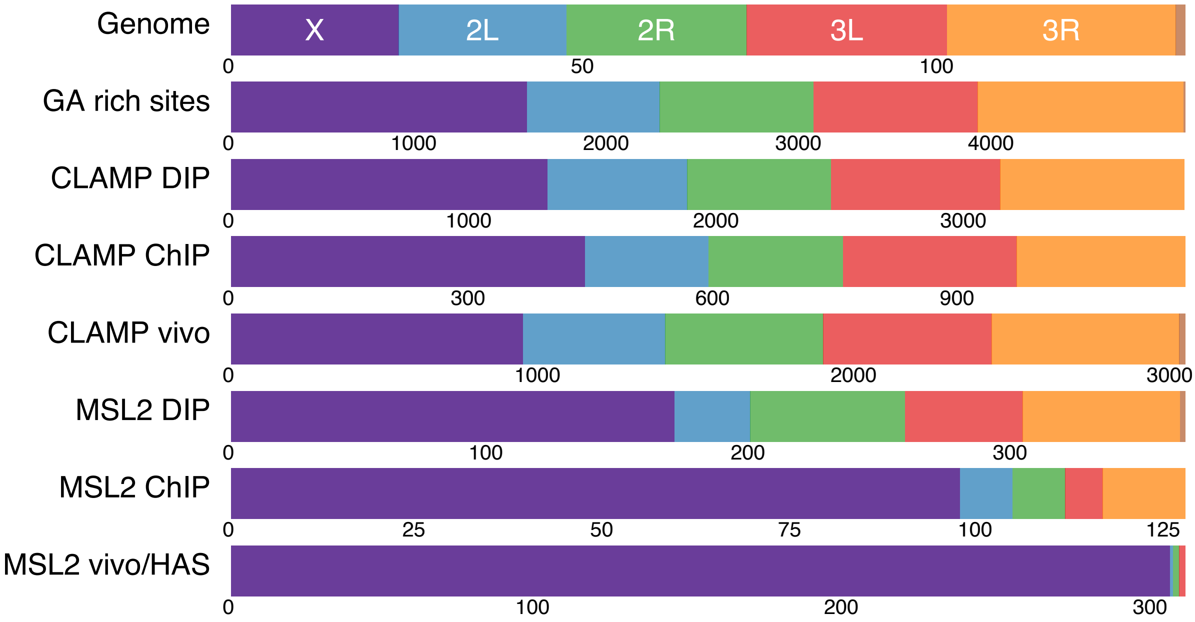


Figure 29: Distribution of peaks as determined by IP. The number of peaks found for each set is annotated underneath and the location of the peaks is color coded. The sizes of the chromosomes are indicated as „genome“ for comparison.

3.3.5 Chromatin Binding is ATP-dependent

ATP-depended nucleosome remodeling is abundant in DREX (Alexiadis et al., 1998; Ito et al., 1997; Tsukiyama and Wu, 1995; Varga-Weisz et al., 1997). Most transcription factors are dependent on this remodeling to open up windows of opportunity in which their binding sites are freed from nucleosomes. Only pioneering factors are able to bind their respective binding sties independently of histones. To test if MSL2 or CLAMP have pioneering properties I applied the standard ChIP-seq protocol after chromatin assembly in vitro with a single difference. After 4 hours of chromatin assembly ATP was depleted using hexokinase and glucose (Tatei et al., 1989). This abolished all DNA binding so that the amount of DNA retrieved did not allow for sequencing library preparation. It seems that neither MSL2 nor CLAMP can bind their targets in absence of ATP. This is true even in presence of CLAMP indicating that also CLAMP itself is not able to open up MSL2 binding sites in absence of ATP. They depend on nucleosome remodelers to randomly open up specific binding sites before they can interact with them.

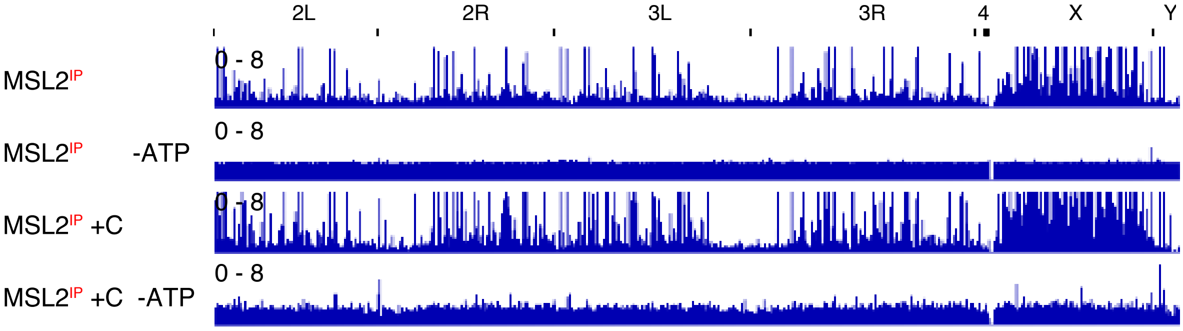


Figure 30: MSL2 binding is ATP-dependent. Browser profile from MSL2 binding in presence or absence of CLAMP and/or ATP as determined by ChIP-seq. Coverage was normalized against the respective controls, the maximum values for each window are shown scaled as indicated. Adapted and reprinted with permission from Oxford University Press (Eggers and Becker, 2021).

3.4 Competition and Cooperation

3.4.1 CLAMP Binding Influences Nucleosome Patterns

CLAMP seemed to bind to chromatin as to DNA in a manner that suggests full transparency of the chromatin to transcription factors. This could be caused either by nucleosome remodelers opening up windows of opportunity for CLAMP to bind, or because CLAMP binding sites are generally nucleosome-free. Nucleosome mapping in the absence of CLAMP showed that its binding sites are actually preferentially occupied by nucleosomes, as shown by the aligned cumulative plots showing nucleosome occupancy at 4041 CLAMP binding sites. Remarkably, as CLAMP is added, nucleosomes are shifted to form phased arrays around the binding site (Figure 31, center). DNA binding proteins can act as boundaries for nucleosome sliding. If they bind tightly enough to not be displaced by the remodelers, nucleosomes will be shifted against them as a border. This leads to phased nucleosomal arrays around such sites. Phaser is such a tight DNA binder whose binding leads to phased nucleosomes next to its target sites and indeed the data shows that these sites are phased independently from CLAMP (Figure 31, left). CLAMP itself can also act as such a boundary. Additionally, CLAMP not only phases nucleosomes around its own binding sites, but also around HAS, leading to a similar pattern at these sites (Figure 31, right). The highly dynamic repositioning of nucleosomes can be explained by the abundance of nucleosome remodelers present in DREX (Längst and Becker, 2001).

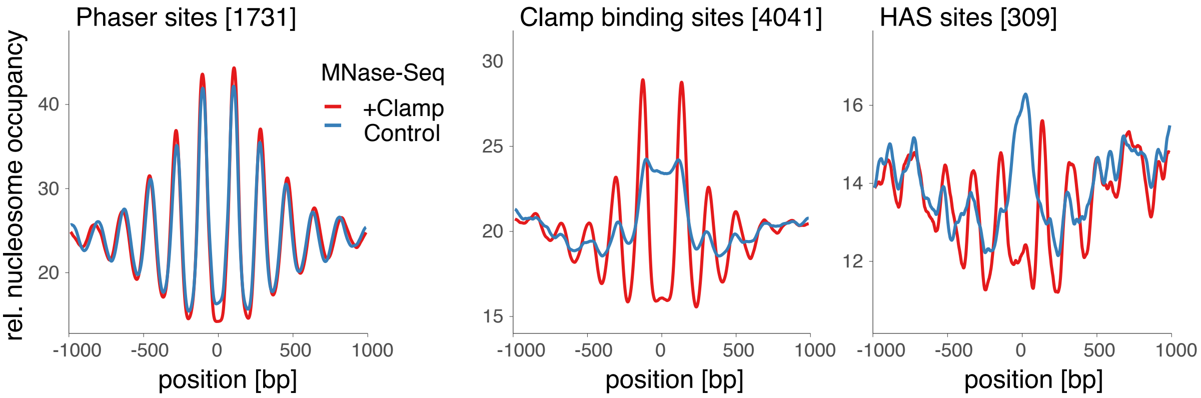


Figure 31: CLAMP phases nucleosomes around its binding sites. Chromatin was assembled in the presence and absence of CLAMP and averaged nucleosome dyad densities were determined by MNase-seq of the mononucleosomal bands. Cummulative nucleosomal occupancy of three biological replicates are shown relative to the motif position at the sites of interest (n=3). Adapted and reprinted with permission form Oxford University Press (Eggers and Becker, 2021).

3.4.2 CLAMP Recruits NURF

The same in vitro ChIP approach as in previous experiments was employed to elucidate how CLAMP binding leads to phased nucleosomal arrays around its binding sites. DREX contains nucleosome remodeling factors that act in an untargeted manner such as CHRAC or ACF (Längst et al., 1999; Varga-Weisz et al., 1997), but also NURF. NURF has been shown to be recruited by a range of transcription factors, including GAF and CLAMP (Judd et al., 2021; Urban et al., 2017a). As DREX already contains intrinsic NURF, chromatin was assembled in presence and absence of CLAMP and then the resulting chromatin was assayed for CLAMP and NURF binding by ChIP-seq and for nucleosome distribution by MNase sequencing. After calculating the average genome coverage for all replicates the binding was plotted at known CLAMP binding sites and Su(Hw) sites as a control, as NURF is also known to be recruited there and these sites also show phased nucleosomal arrays (Bohla et al., 2014). NURF is recruited to CLAMP binding sites only in presence of CLAMP. Specifically, the binding profile shows two peaks around the CLAMP binding sites suggesting that

CLAMP directs NURF to the neighboring nucleosomes (Figure 32, right). At the same time, NURF locates also to Su(Hw) sites in absence of CLAMP. This binding is slightly reduced in presence of CLAMP as CLAMP competes with Su(Hw) for NURF binding and by dilution of the ChIP-seq signal at Su(Hw) sites through additional peaks caused by CLAMP. CLAMP is only rarely recruited to Su(Hw) sites, despite earlier research suggesting CLAMP and Su(Hw) cooperation (Jordan and Larschan, 2021). It is possible that CLAMP promotes NDRs and PNAs around its binding sites through a mechanism involving NURF recruition, but CHRAC and ACF could also be responsible for the observed effects. Further experiments trying to deplete NURF from the extract to confirm this hypothesis unfortunately failed to show any effects on phasing, possibly due to weak antibody binding and insufficient depletion of the intrinsic NURF.

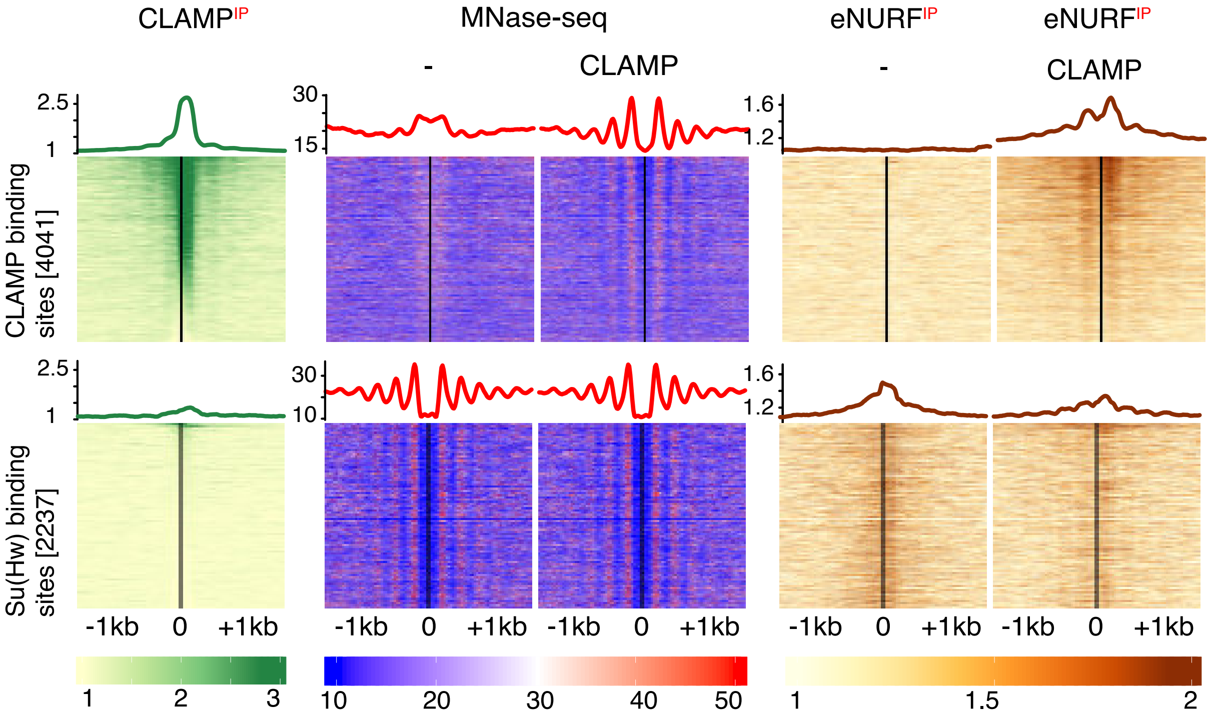


Figure 32: CLAMP recruits NURF around its binding sites. Chromatin was assembled in the presence and absence of CLAMP. For the IP the signal was normalized to the control at CLAMP and Su(Hw) binding sites and coverage windows of 2000 bp around the sites were cut out, aligned and the mean for each column calculated. Data was illustrated by average profiles (top) and heatmaps of individual regions (bottom). The profiles summarize the results of two biological replicates (n=2). All heatmaps were sorted by the binding strength of the first heatmap (CLAMP). Adapted and reprinted with permission from Oxford University Press (Eggers and Becker, 2021).

3.4.3 CLAMP Influences MSL2 Binding

MSL2 and CLAMP can directly interact and support each other's binding in vivo and in vitro (Albig et al., 2019). In a chromatin context this cooperation could lead to improved MSL2 enrichment on the X chromosome. To explore how the presence of CLAMP enhances MSL2 binding, the in vitro MNase sequencing results allow a number of observations. For one, CLAMP addition leads to NDRs and phased nucleosomal arrays not only around its own canonical binding sites but also around HAS (Figure 31, right). HAS are also nucleosome-free in vivo (Albig et al., 2019; Ramirez et al., 2015; Straub et al., 2013). This happens because CLAMP binds to most HAS. This cooperation has been demonstrated before as CLAMP and MSL2 stabilize each other's binding in vivo and vitro at these sites (Albig et al., 2019; Soruco et al., 2013a), but it gives first insights into how CLAMP promotes this additional MSL2 binding. The data also shows that in absence of CLAMP MSL2 HAS are preferentially occupied by nucleosomes, again explaining why MSL2 binding is ATP dependent. Looking at the ChIP-seq profiles of MSL2 allows teasing apart how CLAMP influences MSL2 binding specifically.

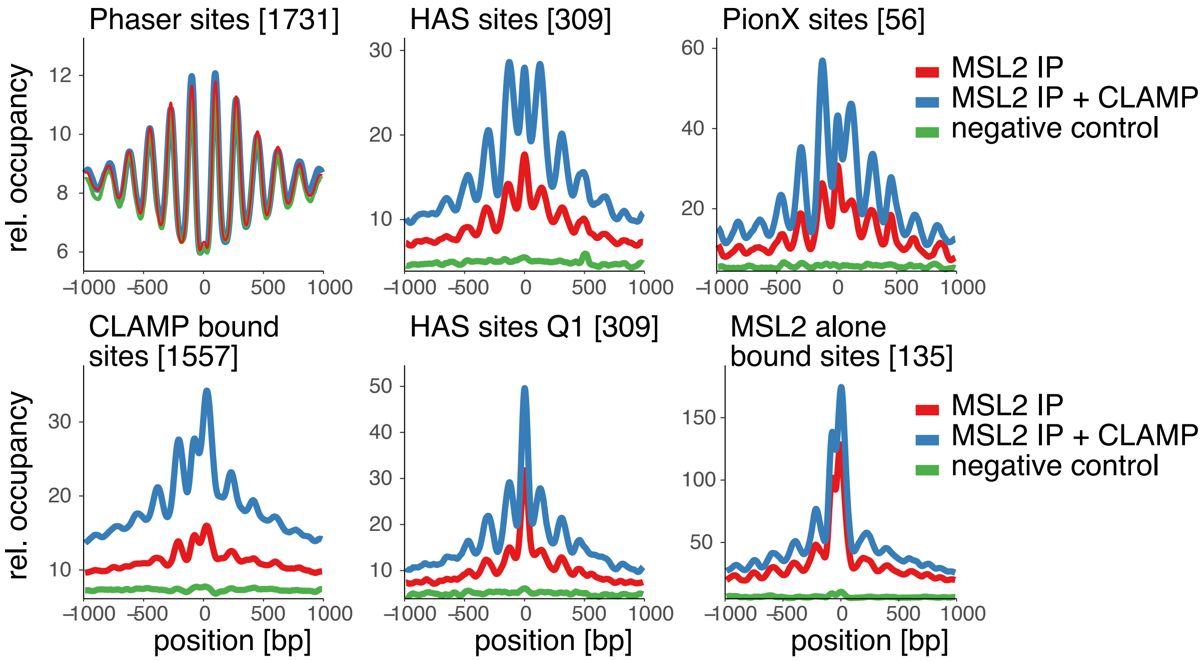


Figure 33: MSL2 binding is enhanced by CLAMP. Chromatin was assembled, then MSL2 and CLAMP were added as indicated and allowed to bind, before samples digestion with MNase and MSL2 ChIP-seq. Signal from all indicated locations was cummulated by cutting out 2000 bp windows around the sites and averaging the genomic coverage. The profiles summarize the results of three biological replicates (n=3). The normalized read coverage is drawn relative to genomic position around the sites. Q1 only considers the small fragments <125 bp of each dataset. Number of sites are annotated in brackets.

First, looking at Phaser sites, which serve as controls for the MSL2 IP, the quality of chromatin assembly can be assessed. The assembly degree of all samples is very similar allowing for a high comparability between them (Figure 33, top left). As seen before, the addition of CLAMP increases MSL2 binding to HAS (Figure 33, top center). Paired-end sequencing allows to separate the reads by fragment size. To pinpoint precise binding of a TF this can be informative, for example CLAMP effects differ for different fragment size. In the Q1 fraction the peaks are narrower and the additional binding through CLAMP is not as pronounced at the nucleosomes adjacent to the main peak (Figure 33, bottom center). Also, after addition of CLAMP the number of binding sites is increased from 131 to 353 while at the same time the X specificity is reduced. This can be explained by looking at sites bound by CLAMP in vitro (Figure 33 bottom left). MSL2 alone binds these sites only weakly while this binding to mostly unrelated sites is strongly promoted by CLAMP. Sites already bound by MSL2 alone (Figure 33 bottom right) do not benefit from the addition of CLAMP. This indicates that these sites are bound irrespective of CLAMP and are already saturated. Addition of CLAMP therefore doesn't enhance binding at all sites equally, but allows MSL2 to bind more and new sites. As CLAMP is only slightly enriched to the X, the overall X specificity of MSL2 is decreased by this cooperation. The percentage of peaks that overlap with HAS also drops from 21% to 16% through addition of CLAMP while the overall number of bound HAS increases (Figure 34). This mirrors the data seen in DIP experiments where CLAMP enhances MSL2 binding but mostly reroutes MSL2 to CLAMP binding sites. In vivo, on the other hand, CLAMP interaction was shown to only occur at HAS (Albig et al., 2019).

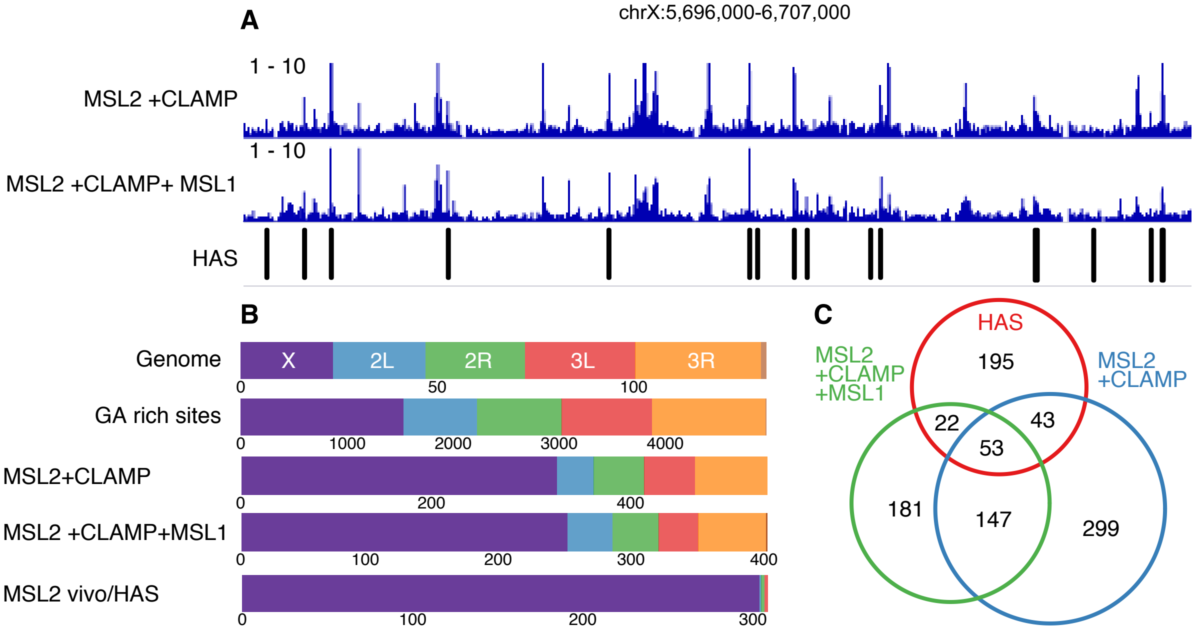


Figure 34: MSL1 has no significant impact on MSL2 specificity in vitro. (A) Browser profile from MSL2 binding in presence of CLAMP and +/- MSL1 at a representative locus determined by ChIP-seq. Coverage was normalized against the respective controls, the maximum values for each window are shown scaled as indicated and HAS are annotated. (B) Distribution of peaks as determined by IP. The number of peaks found for each set is annotated underneath and the location of the peaks is color coded. The sizes of the chromosomes are indicated as „genome" and the occurrence of GA-rich sites is shown for comparison. (C) Venn diagram depicting the overlap of peaks as determined by ChIP-seq of MSL2 under the different conditions and HAS.

3.4.4 MSL1 has no Influence on MSL2 Binding Specificity in vitro

As the addition of CLAMP was not sufficient to achieve faithful X discrimination by MSL2 it is possible that further assembly of the complex could improve results. MSL2 and MSL1 form a heterotetrameric complex mediated by the MSL2 RING domain and the N-terminal domain of MSL1 (Copps et al., 1998; Hallacli et al., 2012). Indeed, MSL1 is thought to build the backbone of the whole DCC complex. The N-terminus of MSL1 is known to promote X chromosome binding, self-association and MSL2 binding (Li et al., 2005a). It is therefore possible that MSL1-MSL2 interactions in vitro would change the DNA binding affinities of MSL2, aiding the overall targeting. Addition of MSL1 to the standard ChIP protocol did not improve MSL2 binding by IP (Figure 34, A,C). As the MSL1 dataset is comprised of only one replicate there are fewer peaks (Figure 34, C) than in the control for technical reasons, but still the profiles look similar (Figure 34, A) . Also, the X-enrichment of MSL2 remains the same even after addition of MSL1 (Figure 34, B). The data suggests that for in vitro reconstituted chromatin MSL1 does not significantly contribute to the targeting of MSL2 either because the MSL1-MSL2 complex has not been formed or because this incorporation does not alter the DNA binding affinity of MSL2 in vitro.

3.4.5 H1 Reduces Overall Binding of MSL2

The addition of CLAMP to the MSL2 IPs led to a lower overall X-chromosomal specificity, not through loss of correct binding but through additional autosomal binding. This indicates an inability of nucleosomes to occlude the decoy sites. The Drosophila preblastoderm extract lacks histone H1 which in vivo binds to almost all nucleosomes and decreases DNA accessibility. Instead, DREX-assembled chromatin is rich in bigH1 and HMG-D which replace H1 during early development (Climent-Canto et al., 2020; Ner and Travers, 1994). Therefore, I explored whether incorporation of H1 would lead to additional occlusion of nonfunctional sites. The necessary protein concentrations to achieve the physiological 1:1 ratio of nucleosomes to H1 were determine by observing the change in linker length by MNase sequencing as described in 3.1.5. A concentration of 100 nM was determined to be sufficient for full occupancy of H1. At this concentration there were no significant changes of nucleosome occupancies (Figure 35, right). Nonetheless, MSL2 IPs in presence of CLAMP returned about 75% fewer peaks if H1 was added to the assemblies. However, the binding of MSL2 was reduced uniformly throughout the genome, therefore did not affect the X-over-autosomal enrichment. H1 seems to reduce accessibility of binding sites in vitro as it does in vivo. This is the first study describing the change of chromatin positioning and TF binding in regard to H1 on a genome-wide level in vitro. Another way to distinguish the peaks found in ChIP-seq is to analyze the predominant motif found within these peaks. To this end, a MEME analysis (see 2.6.4) of the respective peaks in presence and absence of H1 found a slightly shorter and more concise MRE motif to be preferred in the presence of H1. This, however, did not lead to improved targeting of MSL2 in vitro (Figure 36). H1 was consequently excluded from further experimentation as the preblastoderm chromatin lacking H1 is arguably closer to the chromatin state during dosage compensation establishment during early development and H1 did not improve specificity in vitro.

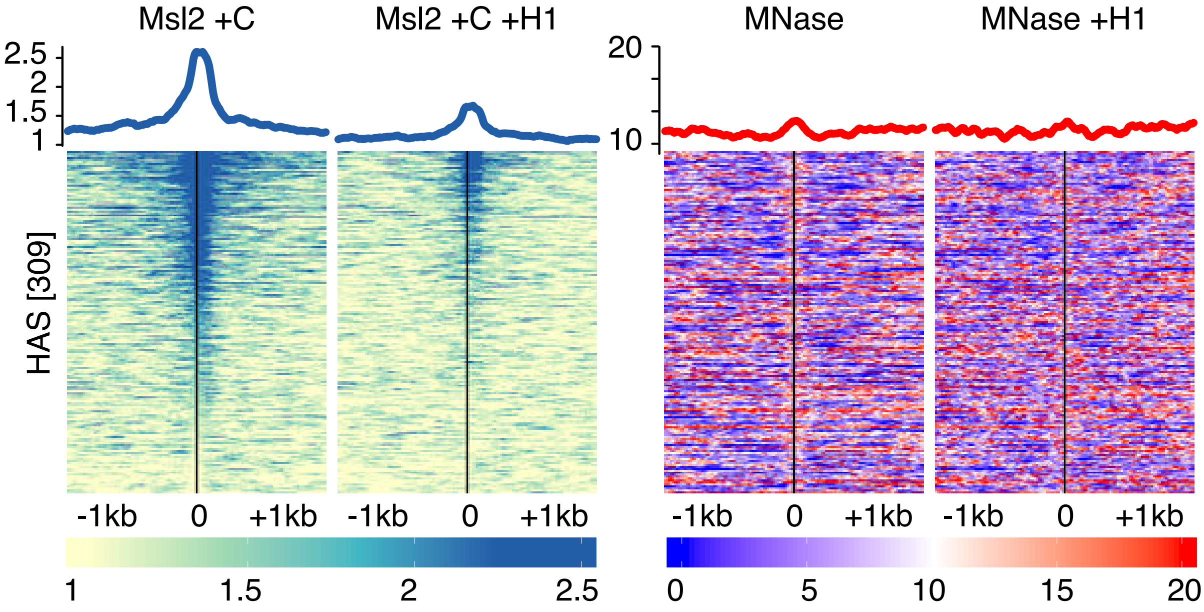


Figure 35: Effects of H1 on MSL2 binding and nucleosome positioning around HAS. Chromatin was assembled in the presence and absence of 100 nM histone H1 and the factors were added as indicated and then digested with MNase. Samples were split and analyzed by MSL2 ChIP sequencing and MNase sequencing. For the IP the signal was normalized to the control at 309 high affinity sites (HAS) and illustrated by average profiles (top) and heatmaps of individual regions (bottom). Coverage windows of 2000 bp around the sites were cut out, aligned and the mean for each column calculated. The profiles summarize the results of 2 biological replicates (n=2). All heatmaps were sorted by the binding strength of the first heatmap (MSL2 IP + CLAMP). Adapted and reprinted with permission from Oxford University Press (Eggers and Becker, 2021).

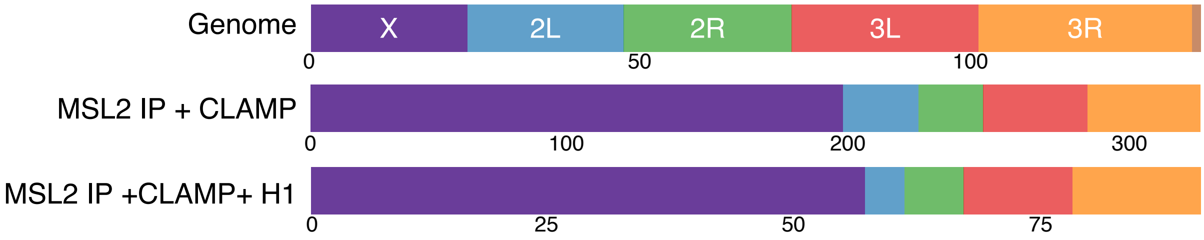


Figure 36: Distribution of peaks as determined by IP. The number of peaks found for each set is annotated underneath and the location of the peaks is color-coded. The sizes of the chromosomes are indicated as „genome“ for comparison.

3.4.6 Shape Analysis of CLAMP-assisted Binding

In presence of CLAMP MSL2 binds to 343 sites with about 60% fidelity to the X chromosome. At this point it remained unclear if the sites bound on the X chromosome and the autosome differ in any way which would allow MSL2 to differentiate between functional und nonfunctional sites in vivo. For this analysis, I assumed that all X chromosomal sites are functional MSL2 targets while the autosomal sites are "decoy" sites by definition. Performing a MEME search to determine the predominantly bound motifs it seems that the PWM of X and autosomal sites are very similar (Figure 37) when it comes to the pure base pair sequence. Consequently, as was shown by IP, MSL2 does not differentiate between these groups in vitro in presence of CLAMP. PWMs show only the probability of each base at each position and don't give any information about certain base pair sequence successions. I applied DNA shape analysis to examine the potential sites to find more subtle features that could characterize the true MSL2 binding sites.

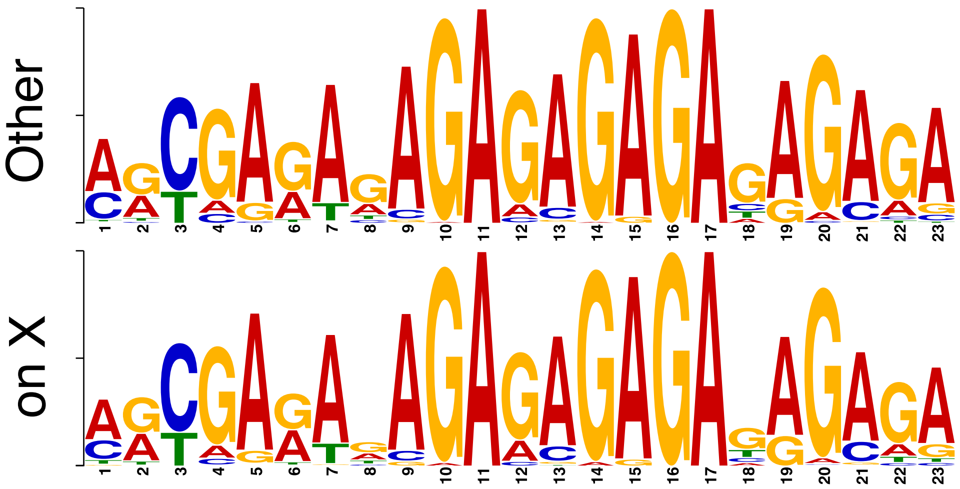


Figure 37: MEME shape analysis of MSL2-bound sites. X chromosomal sites (on X) are very similar to sites bound on the autosomes (Other). This raises the question of how MSL2 can distinguish between them.

For a shape analysis to yield any sensible information, bound motifs must be aligned with base pair precision. As the classical MRE motif with its long GA repeats does not allow for such precision, I used the slightly more directional PionX motif for sequence alignment. MSL2 peaks were sorted according to chromosome and surveyed for the PionX motif. Then all found motifs were aligned and surveyed for their shape features. Then shape feature values were summarized and the median was used for data representation, as shape values are discrete rather than continuous. Also, the data cannot be assumed to be normally distributed. Therefore, I used the Wilcox test to assess whether the populations of 5 different shape features have the same distribution. Looking at Propeller Twist and Roll specifically, 2 regions specifically stand out where the bound motifs on autosomal and X chromosomal sites differ (Figure 38). X-chromosomal sites have a particularly low Propeller Twist and DNA Roll at position 5-7, just before the first GA repeats. The shape features hint that GA dinucleotides are particularly rare between the main motif and the CAC extension of the motif present at the 5' end, but only on X chromosomal sites. The findings indicate that functional sites differ from nonfunctional ones in composition of GA dinucleotides and are more significantly different from a simple GAGA repeat sequence than autosomal sites.

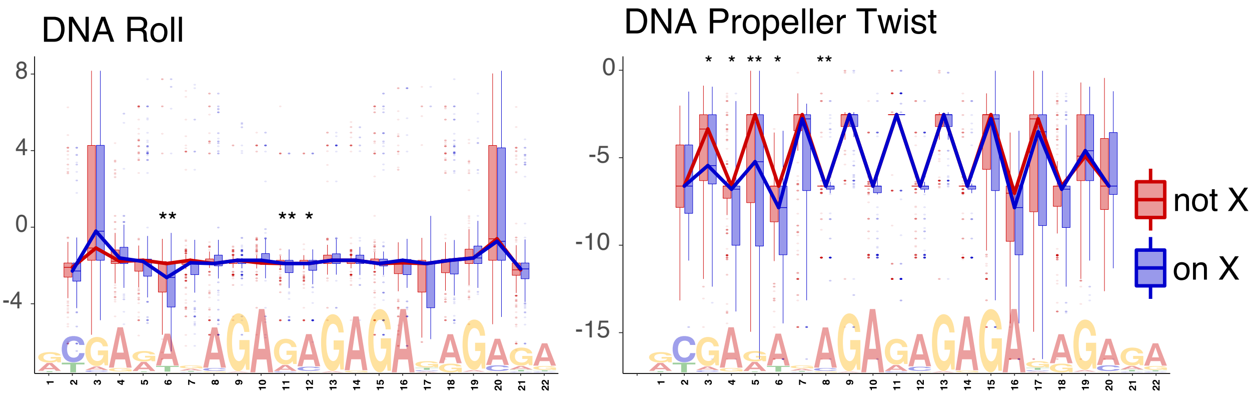


*Figure 38: DNA shape feature predictions. Sites bound by MSL2 were searched for the PionX PWM binding motif by FIMO. If multiple hits were scored for a single peak only the best hit was considered. X chromosomal hits were separated from the autosomal sites and DNA shape features were predicted, as assumably sites on the X are more functional. The consensus sequence is annotated underneath as orientation. The data plotted represents the median at each position. These values have been connected to simplify the visualization. The distribution of values at each position was then tested by a Wilcox test to determine where the two sets differ statistically from each other. Significance is annotated and represented by $p \leq 0.05$ as *, $p \leq 0.01$ as **. This was done for DNA Roll and Propeller Twist. Roll is an interbase feature, for simplicity +1 reflects the Roll from base 0 onto +1.*

This hints at a discriminating factor between the autosomal and X chromosomal sites, which seems not to be determined by the DNA recognition through MSL2 alone, as MSL2 bound indiscriminately to both groups. Additional competition to GA-binding by a known GA binder could hypothetically benefit MSL2 targeting by occlusion of decoy sites. This could allow MSL2 to bind more faithfully to its functional sites. A “TomTom motif comparison tool” database search (Gupta et al., 2007) for known $(GA)_n$-binding proteins in the Drosophila genome revealed GAF as the best candidate.

3.4.7 GAF Binds to in vitro-reconstituted Chromatin

The GAGA factor (GAF) binds GA-rich sites at many regulatory regions. In vivo, GAF recognizes the minimal 5 bp motif GAGAG. It was, however, shown to also oligomerize thereby increasing its affinity to longer GA repeats (Katsani et al., 1999; Omichinski et al., 1997; Wilkins and Lis, 1998). First, I tested if GAF binds to known GAF binding sites (Kaye et al., 2018) using DREX-assembled chromatin as a substrate, on its own and then in presence of MSL2 and CLAMP mimicking the experimental conditions. GAF is not present in DREX (Figure 24)(Tsukiyama et al., 1994; Wall et al., 1995) and can therefore be titrated in as necessary. I used ratios of GAF to CLAMP/MSL2 of 1/3 (8 nM, low), 1/1 (25 nM, equimolar) and 3/1 (75 nM, high) in the binding assays and split the samples to monitor GAF and MSL2 binding simultaneously.

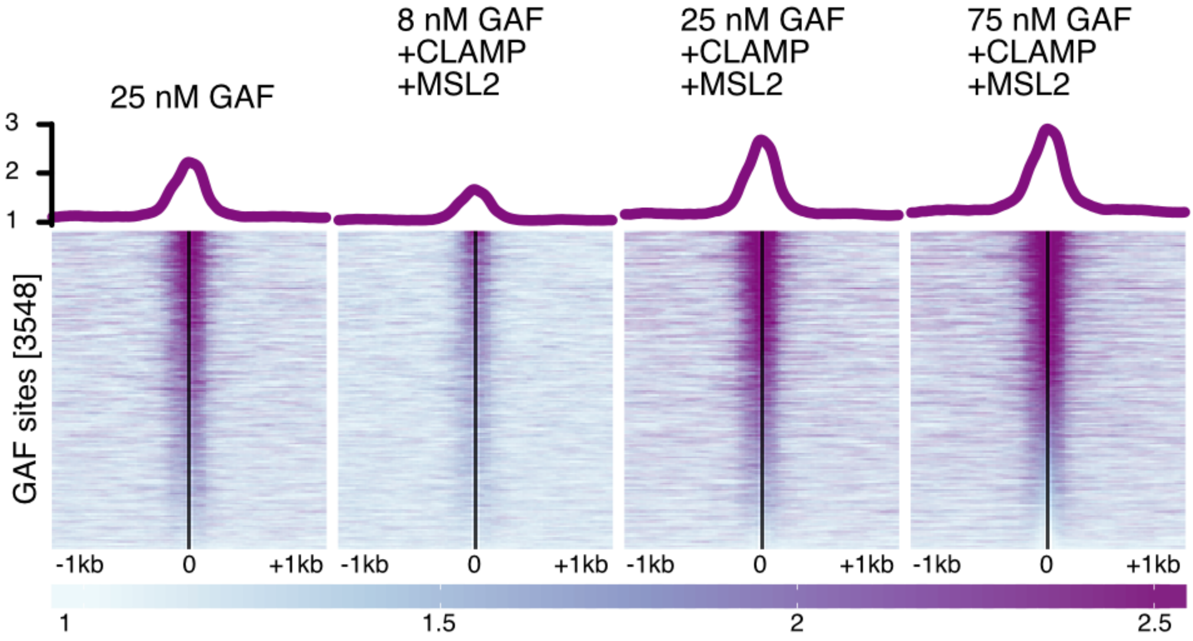


Figure 39: GAF binds the known in vivo GAF sites in vitro. After chromatin assembly proteins were added as annotated and GAF binding was determined by ChIP-seq. Enrichment of GAF was normalized to the negative control at 3548 known GAF binding sites (Kaye et al., 2018) and illustrated by average profiles (top) and heatmaps of individual regions. Coverage windows of 2000 bp around the sites were cut out, aligned and the mean for each column calculated. Two biological replicates were normalized and summarized values are shown. Concentration of MSL2/CLAMP were 25 nM where present. GAF concentrations were 8.3, 25 and 75 nM as indicated. Heatmaps are sorted by the signal strength of the 75 nM GAF IP in presence of MSL2/CLAMP in a 100 bp window around the center. Adapted and reprinted with permission from Oxford University Press (Eggers and Becker, 2021).

The binding of GAF to each known binding site was plotted as a heatmap in Figure 39 and the mean peak height from all sites was plotted above the map. Adding more GAF leads to increased occupancy of sites, which in turn allows to modulate GAFs competing influence in the experimental system by titration. While the strongest binding sites are saturated already with 25 nM of GAF, many more sites are bound in the genome if higher amounts of GAF are added. The genome-wide distribution of GAF shows a stark increase in the number of bound sites suggesting that the system itself is not saturated, only the sites of highest affinity are (Figure 40).

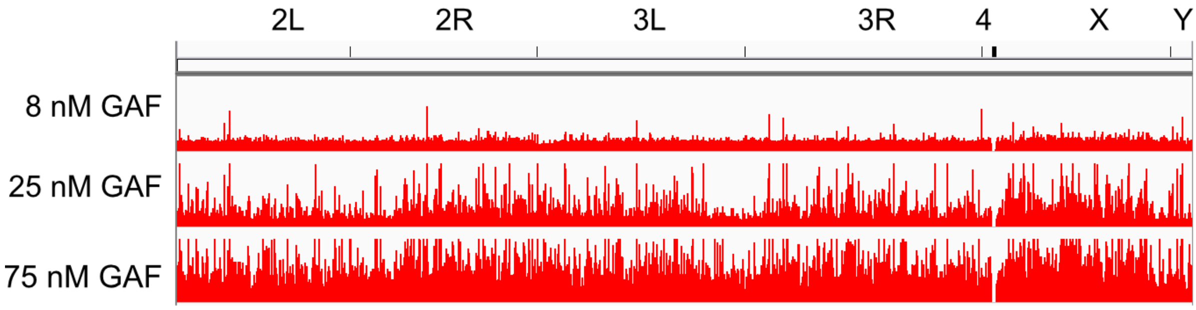


Figure 40: Genomic binding profile of GAF. After chromatin assembly, GAF was added as annotated and its binding determined by ChIP-seq. Signal was normalized to the negative control and represents the mean of 2 replicates. Coverage is shown for the whole genome and chromosomes are annotated on top.

To see if GAF would compete with MSL2 or CLAMP, GAF occupancy in presence of CLAMP and MSL2 was plotted at its own binding sites, CLAMP binding sites and HAS. GAF binds its own binding sites best, followed by CLAMP sites and then HAS. It doesn't strongly enrich at Su(Hw) sites, which serve here as a negative control (Figure 41).

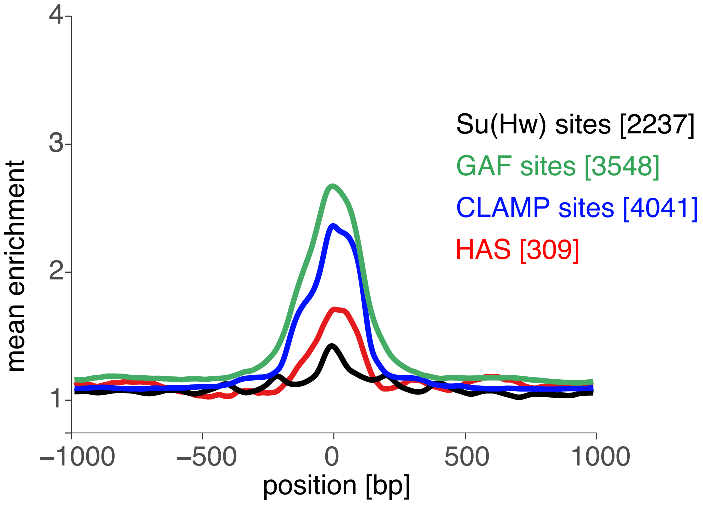


Figure 41: GAF binding to Su(Hw) sites (black), in vivo GAF sites (green), CLAMP sites (blue) and HAS (red). GAF binding was determined by ChIP-seq, normalized to the control and cummulated over all sites of interest. Adapted and reprinted with permission from Oxford University Press (Eggers and Becker, 2021).

3.4.8 GAF Competes with MSL2 and CLAMP for GA-rich Binding Sites

After establishing that GAF binds to its known target sites even in presence of MSL2 and or CLAMP, effects of GAF on MSL2 targeting was quantified. MSL2 in vitro is rerouted from its proper binding sites to mostly unspecific GA-rich sequences by CLAMP (Albig et al., 2019). As these sequences are enriched only slightly on the X chromosome this interaction of MSL2 and CLAMP is unproductive in terms of enrichment. As GAF binds to a similar set of sites and doesn't colocalize with CLAMP it was conceivable that it would outcompete CLAMP from some of the non-physiological sites. GAF recognizes a minimal motif of GAGAG but multiple factors can cover larger repeats.

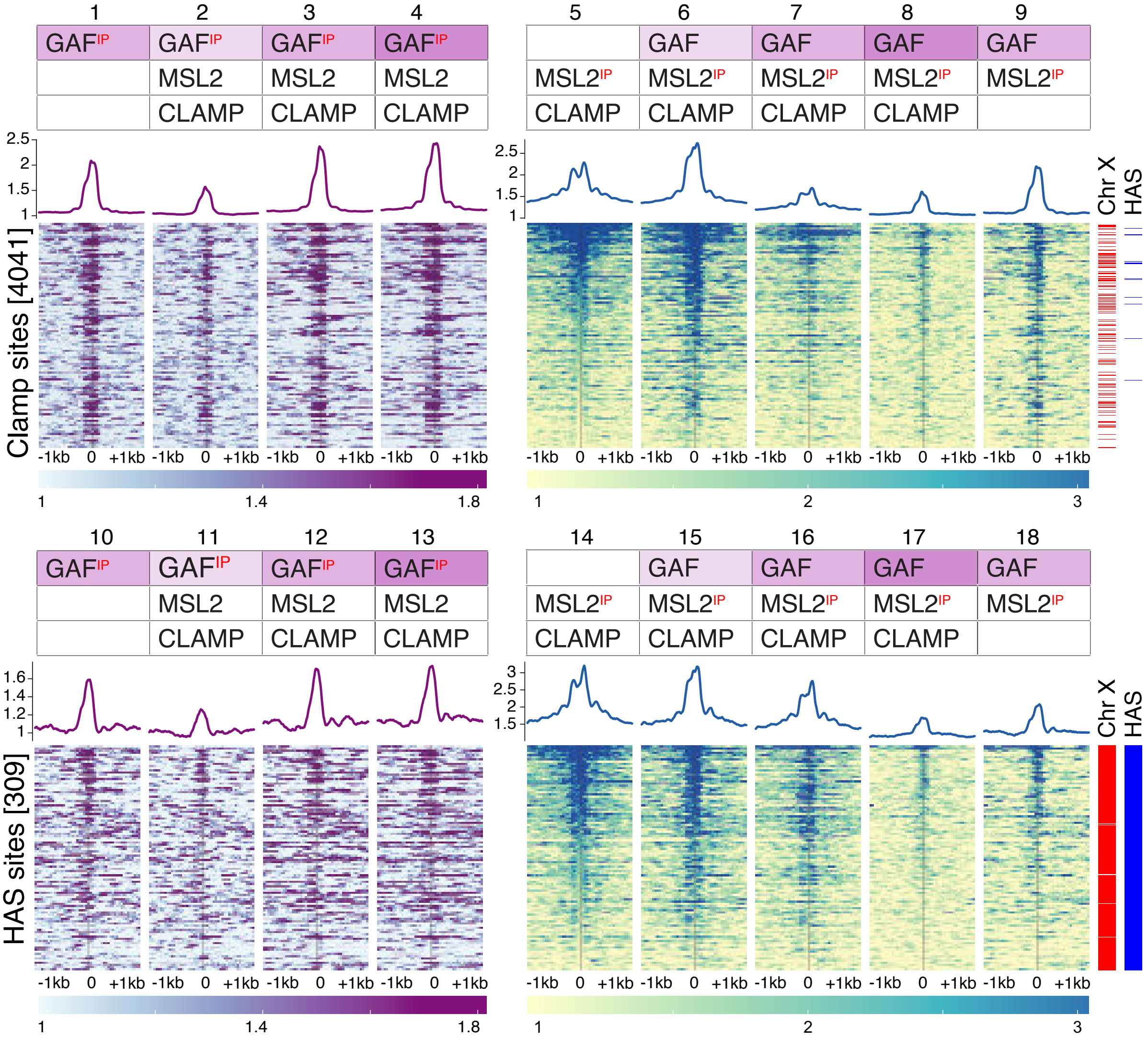


Figure 42: GAF reduces MSL2 binding at nonfunctional sites first. After chromatin assembly proteins were added as annotated and GAF or MSL2 binding was determined by ChIP-seq as indicated by IP. Signal was normalized to the control and summarized from two replicates. Enrichment of the factors at 309 HAS and 4041 CLAMP binding sites are illustrated by heatmaps and average profiles, by cutting out the coverage data of 2000 bp surrounding the binding site. MSL2 and CLAMP concentrations are 25 nM, while GAF concentrations are color-coded and are 8, 25 and 75 nM respectively. Heatmaps were sorted by the enrichment of MSL2 in presence of CLAMP and absence of GAF (5 and 14). Adapted and reprinted with permission from Oxford University Press (Eggers and Becker, 2021).

GAF competition, therefore, should be stronger the more GA repeats are present at a site. Recombinant GAF was titrated into the standard MSL2 ChIP assay to test whether it could occlude GA-rich decoy sites in vitro from CLAMP and MSL2 leading to a higher X-enrichment of MSL2. GAF was used in 1/3, 1/1 and 3/1 ratios to the 25 nM concentration of MSL2 and CLAMP, which was kept constant. Samples were prepared as described in 2.5.6 and split for IP with MSL2- or GAF-specific antibodies. Surprisingly, small amounts of GAF led to an increase of MSL2 binding sites and a decrease in X-enrichment (Figure 42, Figure 44). This can be explained at least partially by nucleosome-mediated cooperativity, which is a general phenomenon for TFs with similar binding motifs or close-by binding sites (Mirny, 2010). As GAF is added it can bind to some GA-rich sites keeping them temporarily accessible for other factors. MSL2, CLAMP and GAF profit from cooperative competition against the occluding nucleosomes enhancing each other's binding unspecifically. Looking at overall MSL2 binding there is a distinct change in MSL2 binding patterns upon the addition of GAF. To be able to compare two groups of GA-rich sites, the binding of MSL2 to functional (HAS) and nonfunctional (CLAMP sites) is shown. Comparing the standard binding reaction (Figure 42, 5 and 14) with MSL2 binding in the presence of 8 nM GAF, there is an increase in binding at HAS and nonfunctional binding sites (Figure 42, 6,15). This overall increase reduces the X-enrichment (Figure 44).

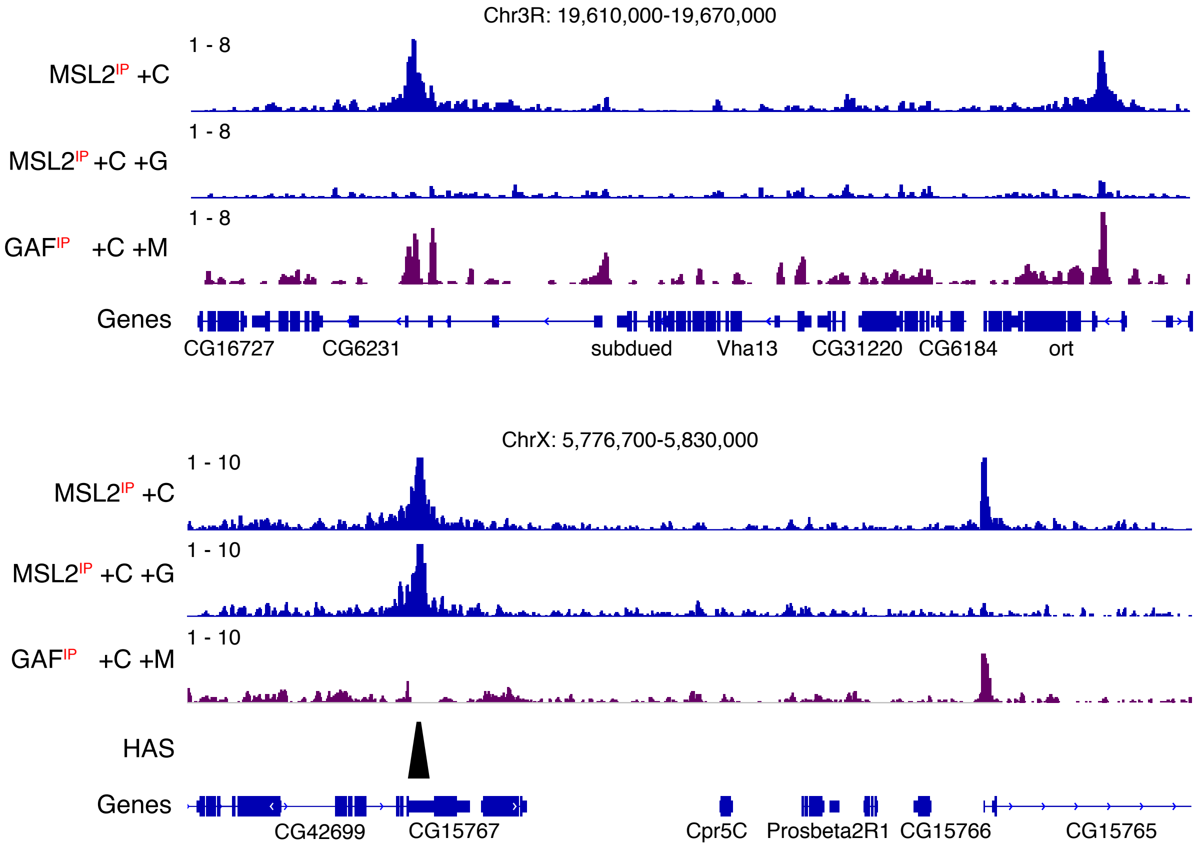


Figure 43: Genomic coverage of MSL2 and GAF in presence of each other. After chromatin assembly proteins were added as annotated and GAF or MSL2 binding was determined by ChIP-seq as indicated by IP. Signal was normalized to the negative control and summarized from 2 replicates. Coverage is shown at representative loci on the X chromosome (bottom) and the autosome Chr3R (top). HAS are annotated and represent functional binding sites. Adapted and reprinted with permission from Oxford University Press (Eggers and Becker, 2021).

At equimolar amounts of GAF, however, GAF competition already strongly reduces MSL2 binding at nonfunctional sites, but not at HAS (Figure 42, 7 vs 16). This trend progresses until at 75 nM of GAF MSL2 binding to nonfunctional and HAS is equally reduced (Figure 42, 8 and 17). At this concentration, GAF seems to effectively outcompete MSL2 also from HAS. Looking at samples in absence of CLAMP, MSL2 IPs show increased binding in the presence of GAF. This increase in binding is accompanied by a strong reduction in specificity, stronger even than the reduction in specificity by CLAMP (Figure 44). GAF interaction is qualitatively different from the CLAMP-MSL2 interaction, which interact directly (Albig et al., 2019). Comparing MSL2 IPs in the presence of CLAMP or GAF respectively (Figure 42, 5 vs 9) it shows that MSL2 selects a different subset of sites in the presence of CLAMP, likely because the CLAMP-MSL2 complex has a different affinity. Although there are additional sites bound the enrichment of MSL2 without CLAMP is reduced in the presence of GAF (Figure 44).

The effects of GAF on MSL2 binding are best visualized by looking at specific representative loci on both X chromosome and autosome (Figure 43). In presence of equimolar amounts of MSL2, CLAMP and in absence of GAF, MSL2 binds to sites on the X and autosome. Also, it binds to some non-HAS sites on the X chromosome. After addition of GAF, the data shows that wherever strong GAF binding is detected, MSL2 binding is now absent. This can be seen specifically well for the two peaks on the autosome and the single non-HAS site on the X chromosome in Figure 43. Remarkably binding of GAF is almost absent from HAS despite their obvious GA-richness.

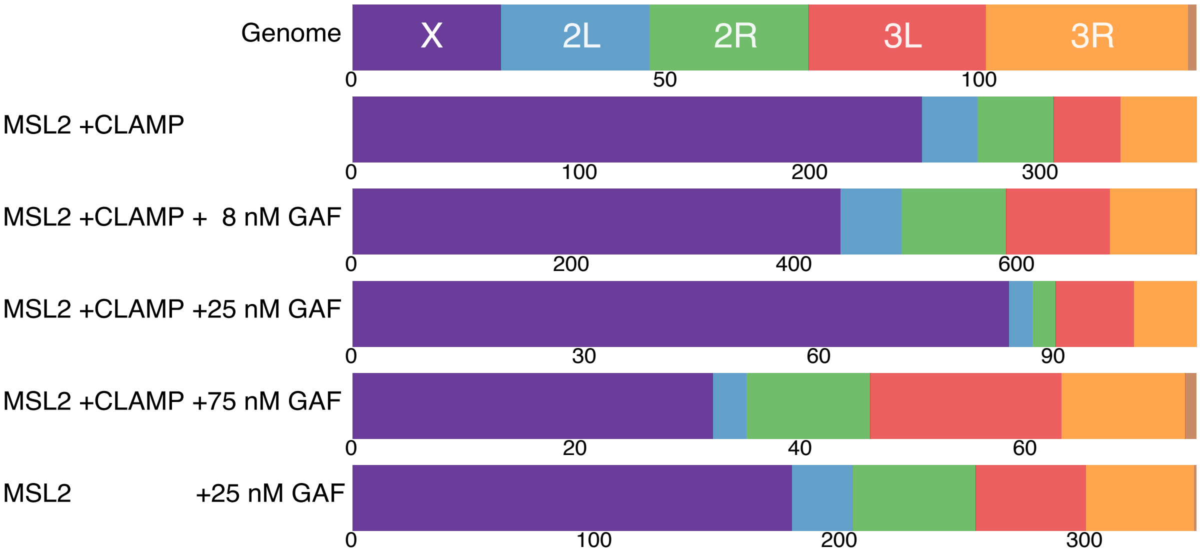


Figure 44: MSL2 enrichment is influenced by GAF competition. Distribution of peaks on the chromosomes as determined by IP. The number of peaks found for each ChIP experiment is annotated underneath and the location of the peaks is color-coded. The relative sizes of the chromosomes are indicated as „genome“ for comparison.

Upon addition of equimolar concentrations of GAF, MSL2 binds fewer sites, which are, however, more enriched on the X. This effect contrasts the global repression observed by H1 competition. It seems that there is a “sweet spot” of GAF competition, in which GAF effectively outcompetes MSL2/CLAMP from nonfunctional sites, while not influencing interactions with the HAS much. The sites now bound by MSL2 are remarkably resistant against GAF competition despite their obvious GA-richness. A MEME analysis of MSL2-bound sites under

GAF competition yielded a motif with the characteristic 5’ extension of the MRE, typical for PionX sites. The sites at which GAF outcompeted MSL2, however, showed extended long GA repeats. On the one hand this demonstrates that GAF competition contributes to MSL2 targeting, while on the other, the question arises as to why some sites are so resistant against GAF binding.

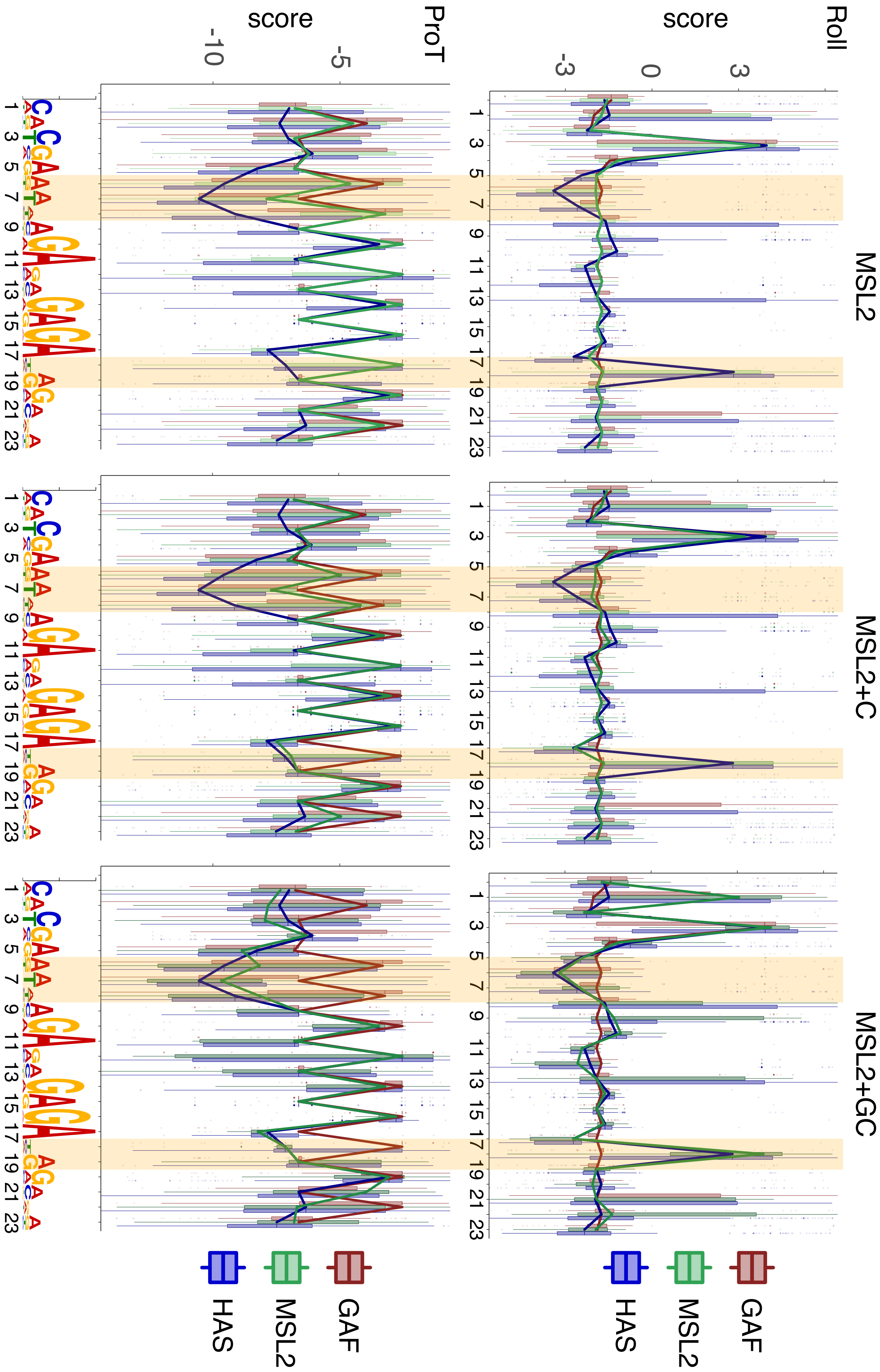
3.4.9 The DNA Shape Influences GAF Specificity.

To elucidate how GAF competition only affects nonfunctional sites while leaving HAS available for MSL2 to bind I considered the distinctive shape features found in the bound versus unbound motifs, mindful of the fact that the PionX signature is also defined by a combination of DNA sequence and shape (Villa et al., 2016). A search of this motif in the genome reveals that of the 500 best hits only about 40% are actually located on the X chromosome and only 71 of them are HAS.

To refine this signature an elaborate bioinformatical analysis was implemented. A requirement for a shape analysis is that all of the motifs under consideration are aligned at base pair precision. As typical binding peaks as determined by ChIP-seq are 150 bp broad and have no strand specificity this necessitates searching for the presence of a defined motif. As previously in 3.4.6, a FIMO search using the PionX PWM was performed as described in 2.6.10 using generous settings to match at least one hit per peak within most of the peaks. Long GA stretches will match multiple times with overlapping sequences, which would impede analysis, so only the sequence with the best fit to the PionX PWM within each peak was considered.

Aligning all bound motifs for each peak set the shape features of all sites were calculated and the values summarized. The “Roll” and “Propeller Twist” are known to be the most influential to transcription factor binding (Gordan et al., 2013; Rohs et al., 2009) and showed the most significant difference when comparing autosomal and X-chromosomal sites bound by MSL2. Therefore, these shapes were plotted along the PionX motif (Figure 45). All values are represented as boxplots and their respective medians were connected for better visibility.

Figure 45: DNA shape features of factor binding sites. A large collection of sites containing PionX PWM sequences were interrogated for MSL2 binding by FIMO (--qv-thresh --thresh 0.01). If an MSL2 peak had multiple PionX motifs only the best-scoring one was considered. For the subset of MSL2-bound motifs the DNA shape feature ‘Roll’ was determined along the 23 base pairs of the motif and plotted relative to the consensus sequence (Roll is an inter-base feature, for simplicity +1 Roll reflects the Roll from base +1 onto +2). This was done for MSL2 alone (n=131), or MSL2 in presence of CLAMP (+C, n=383), or MSL2 in presence of CLAMP and GAF (+C+G, n=111), as indicated in the panels. The data are shown in box plots for each base. The median values were connected with line plot for better visualization. The shape signatures of HAS and of sites bound by GAF serve as references for functional and nonfunctional binding sites, respectively. (D) As in (C) but for the shape feature ‘Propeller Twist’. Reprinted with permission from Oxford University Press (Eggers and Becker, 2021).


Roll
ProT
score
score
3
0
-3
-5
-10
MSL2
MSL2+C
MSL2+GC
1 3 5 7 9 11 13 15 17 19 21 23
GAF
MSL2
HAS

Additionally, the motifs bound by MSL2 under the different conditions were plotted as annotated in each column. MSL2 alone and in presence of CLAMP binds a shape profile similar to what is seen for GAF, while it clearly differs from the canonical HAS. This is obvious already by looking at the median values alone. Interestingly, there is a high variance of shape features for MSL2 at specific sites, such as Roll and ProT at +18 (Figure 45), while the variance for GAF-bound motifs at these sites is very small. This indicates that GAF is restricted to a certain profile while MSL2 is not. The median suggests that the majority of MSL2 sites follow the GAF shape profile. The MSL2 shape features bound in the presence of GAF and CLAMP, however, strongly differ from the profiles seen before and instead highly resemble the characteristics of HAS (Figure 45, right). Especially obvious are the high Roll at position +3 and +18 and a region of low Propeller Twist between +6 an +8. Additionally, and only under these conditions the high +1 Roll signature previously known to be a hallmark of the PionX signature becomes the dominant shape seen at this position in MSL2-bound motifs. The HAS sites show a high variability at this position, as they include PionX and non-PionX sites. The shape profile also shows regions of particularly high variability for HAS motifs and MSL2-bound motifs in presence of GAF and CLAMP. These include positions directly neighboring the central GA repeat and indicate that the uniform GA sequence is interrupted. This becomes most obvious at the DNA Roll at positions +8, +13, and +18 where there is a high variability in functional but not nonfunctional sites. This is particularly interesting regarding the fact that GAF binds a minimal binding motif of “GAGAG”, which seems to be especially selected against at the HAS. Comparing the shape motifs bound by MSL2 alone against the shape motifs bound in presence of GAF it becomes obvious that MSL2 does not distinguish between them as it can bind a broad variety of features at these positions. GAF, on the other hand, seems to bind specifically to the GA repeats. Especially interesting in this regard are also positions with little sequence information, i.e., none of the bases are prevalent, but which show particularly narrow distribution of shape features. At position +18, for example, the sites bound by MSL2 in presence of GAF have a very low Roll and Propeller Twist with a low variance, while the standard GA repeat shows a high roll here. The low sequence information indicates that the exact base is irrelevant as long as it breaks the GA repeat signature. The same is true for the Propeller Twist between +6 and +8.

The data suggests a new and more complex shape feature that distinguishes functional from nonfunctional sites, which is not recognized by MSL2 itself but becomes meaningful only in the context of GAF competition. In the cell-free system, MSL2 can, therefore, only recapitulate in vivo-like binding, in the presence of CLAMP and GAF. This competition aids MSL2 in proper X/autosomal discrimination and enhances X chromosomal enrichment.

4. Discussion

4.1 Cell-Free Genomics

4.1.1 DREX-assembled Chromatin

To determine how transcription factors distinguish their respective targets from a large pool of similar but unrelated sites, earlier approaches employed in vitro assays using naked DNA to test varying DNA binding proteins through DIP-seq (Gossett and Lieb, 2008; Liu et al., 2005) or Systematic Evolution of Ligands by Exponential Enrichment (SELEX)(Jolma et al., 2010). These experiments can be used to determine intrinsic properties of DNA binding proteins in vitro and such findings have been used to predict potential in vivo binding sites (Guertin et al., 2012). These methods, however, have certain disadvantages, as the sites tested by SELEX are artificially constructed and fail to consider repetitive elements that would be present in a complex genome. This is especially problematic considering transcription factors that are known to bind repetitive sequences under certain circumstances such as MSL2, which binds heterochromatic regions in absence of roX (Figueiredo et al., 2014), and GAF, which binds repetitive elements during mitosis (Platero et al., 1998). DIP-seq approaches, on the other hand, consider the whole genome, but lack the physiological chromatin environment in presence of which transcription factors have evolved. As nucleosome competition is a major factor influencing TF binding (Zhu et al., 2018), using chromatinized DNA substrates to test TF binding is preferable.

To provide TFs with a more physiological binding substrate in vitro, nucleosomal arrays can be assembled through salt gradient dialysis (SGD), which was demonstrated for whole yeast genomes (Krietenstein et al., 2012). Here nucleosomes are positioned according to intrinsic sequence preferences, which differ from the physiological positions especially at promoter and coding regions (Kaplan et al., 2009). Nucleosome positioning by SGD can be improved by addition of purified remodelers or whole cell extracts (Krietenstein et al., 2016), but even then, this system fails to faithfully reconstitute native nucleosome positions. Additionally, the approach of Krietenstein et al. utilizes plasmid libraries containing the yeast genome for assembly, as SGD is more efficient on smaller circular DNA substrates than megabase-sized, linear fragments. This introduces normalization issues, as not all plasmids are concentrated equally within the larger library. On the other hand, this avoids DNA DSB effects (see 3.1.6). Using plasmid libraries for assembly is not feasible for metazoan genomes as they are orders of magnitude larger than the yeast genome and would require extensive libraries.

Fortunately, there is a possibility to assemble physiological chromatin on a genome-wide scale using large genomic DNA fragments as substrate. Extracts from Drosophila preblastoderm embryos have been shown to contain large stocks of maternally deposited nucleosomes, nucleosome chaperones and remodeling factors, necessary for chromatin assembly (Becker and Wu, 1992). Using this extract and an ATP source, DNA can be reconstituted into complex, dynamic chromatin containing hundreds of proteins (Baldi et al., 2018b;

Voelker-Albert et al., 2016). Earlier research has demonstrated the ability to assemble chromatin with a dynamic nature and physiological damage response present on fosmids using DREX (Harpprecht et al., 2019) and a pioneering work also reconstituted the Drosophila genome (Baldi et al., 2018b). As these extracts are purified only from embryos collected up to 90 min after egg laying, the assembled chromatin resembles the developmental stage before the onset of ZGA. Testing TF binding on this substrate is highly appropriate as many TF establish their regulatory networks de novo during this phase. Given a sufficiently large fly population, the preparation of such extracts is time- and cost-effective and the resulting chromatin assembly works robustly.

The complexity of whole-cell extracts is, however, also their greatest weakness, as DREX-assembled chromatin may contain proteins not normally found in the nucleus and unknown factors might influence any performed assay. Not all pathways present in DREX, including the nucleosome assembly, are well defined and redundant protein functions or unspecific effects can complicate analysis. Additionally, preblastoderm embryos represent a mix of different cell cycle states, mostly a shortened S- and M-phase (Farrell and O'Farrell, 2014). Therefore, processes that might occur sequentially in vivo will occur simultaneously or in competition in vitro. This is also and especially true for the dosage compensation machinery which in vivo needs to be reestablished after each replication. Also, DREX experiences end resection and a complex damage response at double-strand breaks (DSB) which might influence TF binding (see 4.1.3).

Despite these shortcomings, I decided to establish a reconstituted system using genomic DNA and DREX to assemble chromatin, which allows to map nucleosomes and TF binding genome-wide in a dynamic, competitive environment. This genomics approach allows to probe each individual binding site and provides statistical power by summation of all genomic binding events. Also, the use of whole genomes avoids the problematic exclusion of repetitive sequences from the substrate, which might influence TF binding (Figueiredo et al., 2014; Platero et al., 1998). These sites might act as “sinks” for excess proteins. However, binding to the repetitive sites themselves remains hidden from standard analysis as reads that cannot be assigned to a single specific region are usually excluded.

Cell extracts are inherently heterogeneous between preparations, but the reconstitution of phased nucleosomal arrays (PNA) in DREX-assembled chromatin allows for a convenient quality control of the assemblies. PNAs are unique to chromatin assembled using whole-cell extracts as opposed to SGD or ACF/NAP1-assisted assembly, as the latter lack necessary factors. A phased array will only arise in presence of remodeling activity and a boundary, constituted by tightly bound proteins, in relation to which nucleosomes can be positioned. DREX-assembled chromatin allows Phaser and Su(Hw) sites to be used as controls for the assembly degree and remodeling activity (Baldi et al., 2018b). In presence of CLAMP, also CLAMP binding sites can act as such boundaries and the shifted nucleosomes at those sites can serve as a control for CLAMP activity. As the background of any ChIP-seq is mostly comprised of fragments, which are protected by nucleosomes from digestion, the nucleosomal pattern can always be determined by bioinformatically selecting mononucleosomal fragments, even if the workflow is not targeted to mononucleosomes, but to other factors. Comparing nucleosomal patterns at these sites is, therefore, always possible and allows for a quick quality control readout of experimental conditions.

4.1.2 Cell-Free Genomics Allow Manipulation of Experimental Conditions.

The cell-free approach permits the manipulation of the concentration of factors influencing experimental conditions, such as EDTA, ATP, or reaction time, but also of proteins difficult to manipulate in vivo, such as H1. H1 is only expressed after the zygotic genome activation (ZGA) and therefore absent in early embryos. DREX-assembled chromatin using these early extracts is consequently free of H1. In its absence, proteins like HMG-D and BigH1 bind to nucleosomes (Ner et al., 2001) in a linker histone-like manner and this leads to a chromatin that resembles the state found in early embryos just before the establishment of dosage compensation (Ner and Travers, 1994). The dynamic nature of DREX-assembled chromatin allows titration of H1 into the chromatin assemblies and subsequent quantification of the effects on nucleosome positioning by MNase-seq. The addition of H1 to DREX-assembled chromatin had similar results as earlier in vivo and in vitro experiments (Lu et al., 2009), confirming again the highly physiological way the system reacts to outside stimuli. The linker length was extended and the accessibility of the underlying chromatin reduced. As some of the nucleosome remodelers present in DREX, such as ISWI can also slide chromatosome this was somewhat expected (Clausell et al., 2009; Maier et al., 2008).

The possibility to assemble entire fly genomes into chromatin with or without H1 offers unique opportunities, as experiments of this kind are impossible in vivo. Histone H1 has a 100 amino acids-long and extremely basic C-terminal tail that occupies some of the linker DNA. The globular domain of H1 binds centrally to the nucleosome at the dyad axis while the tail binds to one of the extruding linker DNA strands (Bednar et al., 2017). MNase-seq in relation to H1 presence permits to probe how nucleosome positions change depending on H1. Here, DREX-assembled chromatin is particularly suitable as it can form PNAs with very defined regular nucleosomal positions, from which the linker lengths can be inferred. The possibility to summarize many genomic locations and binding events allows to quantify even small effects. My preliminary analysis showed that the changes in linker length are not uniform for all linkers relative to the boundary site of the PNAs. This could help answer the question if the histone H1 tails exhibit some kind of directionality, which has so far remained elusive. Careful titration and repetition of the shown experiments, in which also the digestion degree must be controlled meticulously, could help to elucidate if H1 tails tend to be all directed to, or away from, a given border or if a random arrangement pattern prevails. However, the findings could be complicated by the presence of BigH1, which is usually responsible to prevents premature ZGA and has an additional long and negatively charged N-terminal tail (Perez-Montero et al., 2013). DREX-assembled chromatin, therefore, cannot model the difference between H1-bound and unbound states, but only between BigH1-bound and, as BigH1 is replaced by H1, H1-bound states.

4.1.3 DREX-assembled Chromatin and the DNA Damage Response

The genomic DNA used for chromatin reconstitution is provided as linear DNA fragments of about 150 kbp. If exposed to DREX, fragments such as these trigger a DNA damage response similar to DSB on chromosomes in vivo (Harpprecht et al., 2019). This response involves phosphorylation of H2A.V over large chromatin domains, but also the recruitment of DNA end-associated proteins such as the Ku complex and Rrp1. Problematically, it also leads to end resection of DNA at the break sites. This response could potentially perturb the binding of transcription factors. To explore the extent of the problem I assembled chromatin with well-defined restriction cuts. Since the genomic locations of these sites are known, chromatin features such as histone depletion, H2A.V distribution and DNA loss can be visualized in cumulative plots. My findings confirm the loss of DNA at break sites, presumably by a mechanism akin to resection on a genome-wide level as reported before (Harpprecht et al., 2019). Considering this, I decided to base my study on randomly fragmented, roughly 100 kb-sized genomes obtained from standard extraction protocols, as any local effects of each break would be averaged out over the population. However, the damage response status of the reconstituted chromatin must be kept in mind. Presence of H2A.V phosphorylation and damage response factors might be widespread throughout the genome and influence the performed assays.

4.1.4 PPPS in DREX-assembled Chromatin

In vivo chromatin within the nucleus is spatially organized. The chromosomes and proteins facilitating gene expression are heterogeneously distributed and undergo a dynamic self-organization by relatively weak interactions such as folding, looping and scaffolding (Meldi and Brickner, 2011; Rippe, 2007). This self-assembly can be facilitated by simple protein-chromatin interactions, but given the right protein concentrations and conditions also a physical phase-separation (PS) can occur (Banani et al., 2017). There are two main mechanisms through which this transition can be driven: a polymer-polymer PS (PPPS), which will result in a collapsed globule, or a liquid-liquid driven PS (LLPS), which results in a liquid-like droplet. The difference is mainly that PPPS depends on soluble bridging factors that bind a polymer, such as the chromatin fiber, to stabilize the phase separation, while LLPS is driven by the interactions between the soluble factors themselves and is not dependent on such a polymer (Erdel and Rippe, 2018).

DREX-assembled chromatin also compartmentalizes into condensed globules, resulting in heterogeneous distribution of chromatin in vitro. It would be interesting to determine which mechanisms underlies this condensation, because LLPS- and PPPS-driven condensation have different functional implications regarding the dynamics of chromatin. Especially the diffusion dynamics of factors unrelated to the formation of these compartments would be different, as LLPS influences this dynamic while PPPS does not (Erdel and Rippe, 2018). This could have major implications regarding the actual concentration of transcription factors within the chromatin, their binding affinities and diffusion rates. Both mechanisms can lead to the formation of dense chromatin compartments, but will have different responses to changes in component concentrations. PPPS compartments sizes are independent of the concentration of factors if above a certain threshold. They scale as a function of the number of possible binding sites for the bridging factors. LLPS compartments, on the other hand, should scale depending on the concentration of factors and the overall molecular composition (Erdel and Rippe, 2018).

These differences allow to distinguish which mechanism is prevalent within the DREX-assembled chromatin. Digesting DREX-assembled chromatin with MNase after compartmentalization, for example, would answer if LLPS or PPPS is prevalent, as PPPS compartments should dissolve in absence of the polymer chromatin fiber. Another possibility is to fluorescently tag and then modulate the concentration of the driving factor. For PPPS the concentration and therefore fluorescent signal within the compartment should scale with the overall concentration while the size of the droplet would remain stable, while for LLPS the size should increase while concentration within the compartment remains stable (McSwiggen et al., 2019).

It is also possible to measure the recovery after photobleaching to distinguish these mechanisms. Using a factor with a higher diffusion rate than binding rate, LLPS predicts that the diffusion is the limiting factor for recovery, therefore a larger bleached area would influence $t_{1/2}$, while in PPPS the binding rate dominates, so that the size of the bleached area should not influence the recovery speed (McSwiggen et al., 2019). It would also be prudent to fluorescently label MSL2 or other assayed transcription factors to measure if they specifically concentrate within the chromatin compartment. This might help to better estimate the ratio of transcription factor to binding site to quantify how the in vitro situation differs from physiological values.

However, the condensates found in DREX-assembled chromatin do not seem to be perfectly round, especially the larger ones, which in itself is a strong indication for PPPS over LLPS.

4.2 Transcription Factor Assays

4.2.1 Developmental Status of DREX

The DREX is purified from embryos collected up to 90 min after egg laying. During the first two hours of development fly embryos undergo rapid cycles of cell division without transcription. The ZGA only starts around the 2 hours after egg laying, only after which many transcription factors are expressed. In respect to the probed factors this means that neither CLAMP, GAF nor MSL2 are expressed in the extract nor is any DCC activity established. However, Both CLAMP and GAF are already maternally deposited even before ZGA (Gaskill et al., 2021; Rieder et al., 2017), but as the DREX protocol separates the chromatin fraction including any bound factors from the cytoplasmic fraction which is further processed into the extract, the proteins might still be absent from DREX.

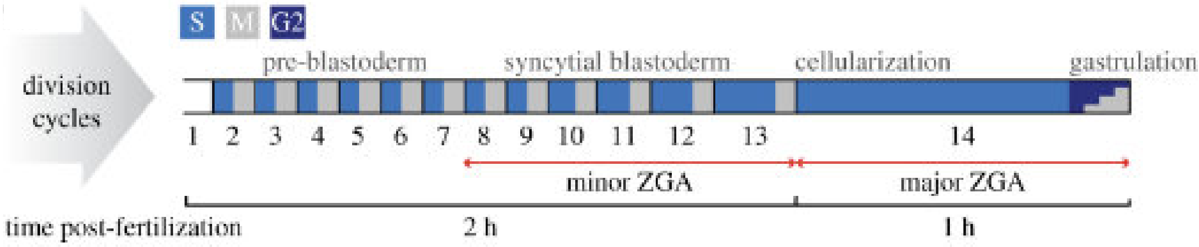


Figure 46: Schematic depiction of the first hours of Drosophila embryo development. The early embryo exists as a syncytium of nuclei undergoing rapid division cycles of repeated DNA replication (S) and mitosis (M). Progressive elongation of S-phase permits time to achieve transcriptional competence from the zygotic genome. The major wave of genome activation occurs at the onset of cycle 14, accompanied by cellularization of nuclei and the introduction of a gap phase (G2). Adapted and reprinted from The Royal Society (creative commons license)(Hamm and Harrison, 2018).

In addition to the maternally deposited proteins, further problems can arise from practical complications. Collection of the feeding plates is done by replacing the collection plates every 90 minutes to ensure no embryo was present on a plate for longer than that time. Unfortunately, this procedure perturbs the flies and they may retain fertilized eggs, which are then laid at an advanced stage in the subsequent collection cycle. Some contamination with older embryos is therefore difficult to avoid. However, most embryos are actually significantly younger than 90 min. The contamination of older embryos as determined by visual inspection made up less than 1% of all collected embryos (personal communication with Peter B. Becker). Additionally, the purification protocol itself leads to exclusion of most of the larger structures, such as membranes, ribosomes and nuclei. To assess the state of DREX, I performed Western Blots assays for the relevant factors and confirmed that the extracts used are free of factor contaminations. Additionally, functional assays showed effects after the addition of factors strongly supporting the assumption that they were not present in significant amounts before. The performed experiments depend on their absence, as it allows to titrate the factors in specifically and distinguish their specific effects.

Adding CLAMP and GAF to this native DREX-assembled preblastoderm chromatin mimics the onset of ZGA as these factors are known to be vital during this developmental stage together with the pioneering factor Zelda (Duan et al., 2021; Gaskill et al., 2021). Most transcription factors are unable to bind to the closed chromatin present before ZGA (Soufi et al., 2015), but CLAMP and GAF are considered pioneering factors in this regard and promote accessibility for other factors by binding GA-rich sites at promotors (Fuda et al., 2015). Levels of GAF, CLAMP and possibly MSL2 are increased at the onset of ZGA, and therefore it is highly appropriate to test MSL2 binding as influenced by these major developmental regulators.

4.2.2 Distinguishing Functional and Nonfunctional Binding

Simply stating that TFs need to select functional from nonfunctional sites is rather obvious, but defining a site as “functional” on the other hand is quite problematic. How can a binding event in an in vitro assay be defined as functional, if there are TFs that bind ambiguous DNA sequences or even sequences that clearly deviate from the consensus (Mulero et al., 2019)? Also, binding sites can only be turned functional through a combination of multiple protein interactions or accumulation of low affinity binding sites (De Kumar and Darland, 2021; Kim and Shendure, 2019). Additionally, the nucleosome positions influence the accessibility of binding motifs, while the 3-dimensional conformation of the chromatin might facilitate cofactor interactions or oligomerization of the TFs, rendering certain sites functional in vivo (Kim and Shendure, 2019; Michael and Thoma, 2021). All these effects can play out differently in different cell types or cell cycle phases, which makes it difficult to determine any binding event as “functional”.

In this problematic situation MSL2 presents a unique opportunity to score the in vitro binding events. As arguably all functional targets of MSL2 are considered to be on the X chromosome, it is not necessary to take all the subtleties into account. While canonical binding sites have been established as HAS, which could be compared to the in vitro profile, the overall X-chromosomal enrichment can also act as a proxy to quantify the truthfulness of binding.

On top of these biological issues, technical challenges have to be considered. ChIP samples are usually crosslinked by formaldehyde to facilitate efficient recovery of TF-bound DNA. Therefore, in addition to the strongly bound sites, some transient binding events might be captured. Depending on the duration of the crosslinking and the factor concentrations these low affinity events might overshadow the stronger ones. Assuming that all sites with high TF affinity are almost continuously bound they will always be captured; additional binding events can only decrease the percentage of functional binding events. Additionally, IP experiments in this regard are inherently binary as they either retrieve a DNA fragment from a given location or not, which might hide the dynamics present in physiological binding events. Adding to this first binary sampling, peak calling represents the next level of non-gradual data interpretation. To simplify the vast amounts of data present in next-gen-sequencing samples, defining binding events is necessary but inherently subjective. Here again a site is either bound or unbound, as defined by arbitrary values, and no subtleties can be attributed. In the case of MSL2 it might even be preferable to just count all reads received from each chromosome and to quantify enrichment of X-chromosomal reads over background, avoiding the error-prone peak calling process altogether.

4.2.3 Successes of in vitro Reconstitution

Despite these problems, in vitro reconstitutions can provide useful answers to long-standing questions. For one, they allow to test the intrinsic binding capabilities of MSL2 and CLAMP to chromatin individually, which was previously tested only on naked DNA (Albig et al., 2019; Villa et al., 2016). It was possible to determine that MSL2 and CLAMP can bind their respective targets within a chromatinized environment, if nucleosomes are dynamically remodeled in the presence of ATP. In absence of ATP, however, no binding could be detected, which argues against the pioneering function, if defined as the ability to bind a target sequence regardless of nucleosome occupancy, as previously suggested for CLAMP (Duan et al., 2021). Rather, CLAMP's pioneering ability seems be connected to its interactions with NURF (Judd et al., 2021; Urban et al., 2017a). This also confirms that DREX-assembled chromatin successfully competes against TFs as previously described for chromatin in vivo (Zhu et al., 2018).

Supporting the pioneering abilities of CLAMP, if acting in concert with NURF activity, the findings suggest that CLAMP can bind the DIP sites even in presence on chromatin. While the binding is reduced overall, the overlap between DIP and ChIP datasets is remarkably high (93% of in vitro ChIP sites are included in the in vitro DIP seq dataset (Figure 28)(Albig et al., 2019)). However, in vivo binding profiles of CLAMP differ greatly and seem to depend on factors not present in vitro, as CLAMP often colocalizes with other factors at non-HAS sites. This was previously demonstrated by ATAC-seq after CLAMP RNAi, which led to reduction in accessibility only at very few CLAMP sites, notably HAS, suggesting that most CLAMP binding sites are kept open also by other factors (Albig et al., 2019).

Conversely, for MSL2 the binding to DREX-assembled chromatin greatly differed from the in vivo and the in vitro DIP situation (Figure 28), suggesting that MSL2, in contrast to CLAMP, cannot bind the intrinsically favorable sites, as determined by DIP, if they are occupied by nucleosomes. This led to an increased X-chromosomal enrichment in the presence of nucleosome competition. The motif found at MSL2-bound sites on DREX-assembled chromatin highly resembles the motif determined in earlier in vivo and in vitro experiments (Alekseyenko et al., 2008). However, some autosomal binding was retained and MSL2 by itself did not distinguish nonfunctional sites with long GA stretches from the functional ones usually harboring shorter GA repeats, hinting that its own intrinsic binding domain is rather promiscuous. Interestingly the MSL2 binding sites, obtained from in vitro ChIP, included motifs with the same shape profiles as those found with free DNA. Considering that wrapping DNA around nucleosomes might distort the double-helix in numerous ways it also wasn't obvious that subtle shape features determined from binding sites of MSL2 in vivo and in vitro would be penetrant in the chromatinized environment. However, the shape analysis of MSL2 bound sites has come up with some of the same basic elements present in HAS (e.g. high roll at base +3) demonstrating that at least at bound sites the DNA is unwrapped insofar that DNA shape features can be read out. This is in line with earlier observations that bound HAS are nucleosome free (Albig et al., 2019; Straub et al., 2013).

Additionally, in vitro reconstitutions allow to test the influences of certain cooperating or competing factors on the binding specificity of another factor. Here we utilized MSL2 as a readout, as its binding profile can easily be quantified in terms of functional vs nonfunctional binding. CLAMP has been implicated in aiding MSL2 binding in vivo and in vitro before (Albig et al., 2019; Rieder et al., 2019; Soruco et al., 2013b; Urban et al., 2017a). Reciprocal increase of binding for MSL2 and CLAMP in the presence of its partner was also seen on DREX-assembled chromatin, recapitulating these findings. CLAMP's effects are mostly thought to originate from its ability to increase chromatin accessibility around its own and its partners binding sites through recruitment of nucleosome remodelers (Urban et al., 2017a). My NURF IPs on DREX-assembled chromatin in presence and absence of CLAMP support these earlier findings. As CLAMP is an essential protein it cannot be deleted in vivo, but one has to depend on the efficacy of partial reduction of the protein levels through RNAi knockdown or similar treatments. In vitro it is possible to simply exclude CLAMP from the experiment. The cell-free genomics approach allows for convenient comparison of CLAMP-dependency of each individual binding event. MSL2 overall binding is enhanced by CLAMP while specificity is reduced. From the graphs in Figure 33 it is clear that addition of CLAMP does not enhance binding at sites already bound by MSL2 by itself, but opens up additional sites for MSL2 binding. MSL2 mostly gains binding at CLAMP binding sites, but as CLAMP is only slightly enriched on the X, promiscuous interaction leads to an overall reduction in X-specificity of MSL2 binding, as seen before (Albig et al., 2019). This interaction also led to increased binding of CLAMP. This reciprocal enhancement of binding is a widespread mechanism of physically interacting factors (Jolma et al., 2015; Morgunova and Taipale, 2017).

GAF is an abundant DNA binding protein occupying many GA-rich sites throughout the genome. It rarely colocalizes with CLAMP in vivo (Albig et al., 2019; Kaye et al., 2018), indicating a strong competition between these two proteins. Similar to CLAMP, GAF is a pioneering protein able to open up chromatin by interacting with NURF (Judd et al., 2021), regardless of nucleosome occupancy. Any TF with a binding site close to a GAF binding sites will profit from these pioneering effects (Granok et al., 1995). Consequently, additional MSL2 binding can be observed at low GAF concentrations before, at higher concentrations, it can outcompete MSL2 from binding. MSL2 benefits from cooperativity with any GA-binding protein, but the cooperativity between MSL2 and CLAMP or GAF respectively is clearly different (Figure 42, panel 9 vs 5). GAF cooperativity is indirect and nucleosome-mediated and decreases specificity massively, while CLAMP cooperativity is direct, leading to a higher affinity for functional sites and a less pronounced reduction in specificity, even though GAF and CLAMP themselves are similarly enriched to the X (Figure 47).

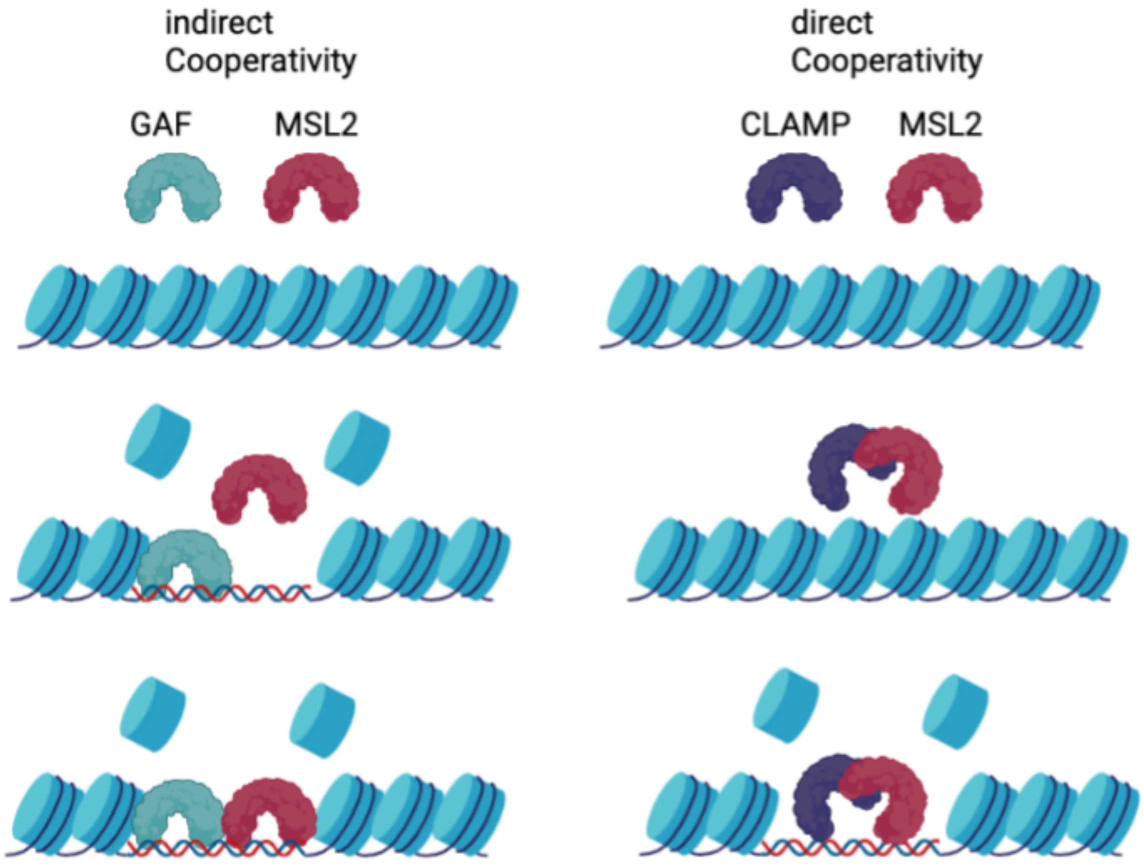


Figure 47: Different modes of cooperativity. MSL2 binding is enhanced by the presence of either GAF or CLAMP but through different mechanisms leading to different outcomes. CLAMP and MSL2 interact directly, while GAF-MSL2 cooperativity is indirect and mediated by nucleosome competition. At low concentrations of GAF, its presence therefore leads to enhanced MSL2 binding. Created with BioRender.com.

GAF cooperativity was only seen for low GAF concentrations, while at higher amounts the additional competition outweighed the cooperative effects and reduced MSL2 overall binding. This helps to explain earlier observations where GAF depletions led to some male-specific lethality and additional erroneous autosomal binding of the MSL2 protein in vivo (Greenberg et al., 2004). This indicates that GAF competition plays a role in DCC targeting even though the effects in vivo were rather small. RNAi experiments reducing the amount of GAF in cells on the other hand, did not effect MSL2 targeting (Albig et al., 2019). However, this experiment depleted GAF after DCC binding was already established. As the DCC is rather immobile after it has bound a HAS (Straub et al., 2005), it is unsurprising that any effects GAF could have on the initial MSL2 targeting are diminished in comparison to its effects after binding is established. Adding GAF during the establishment of binding, as reconstituted in the DREX-assembled chromatin system, shows that GAF easily outcompetes CLAMP and MSL2 from sites containing its GAGAG recognition element. Surprisingly GAF competition was significantly weaker at HAS even though these sites harbor GAGAG elements according to PWM analyses. This posed the question of why GAF competes less with MSL2 at functional sites than at decoy sites. Here, shape features have to be taken into account. Villa et al. showed that many sites that conform to the PionX sequence are still nonfunctional and DNA shape features were used before to improve binding site predictions of TFs in vivo (Li et al., 2017; Mathelier et al., 2016).

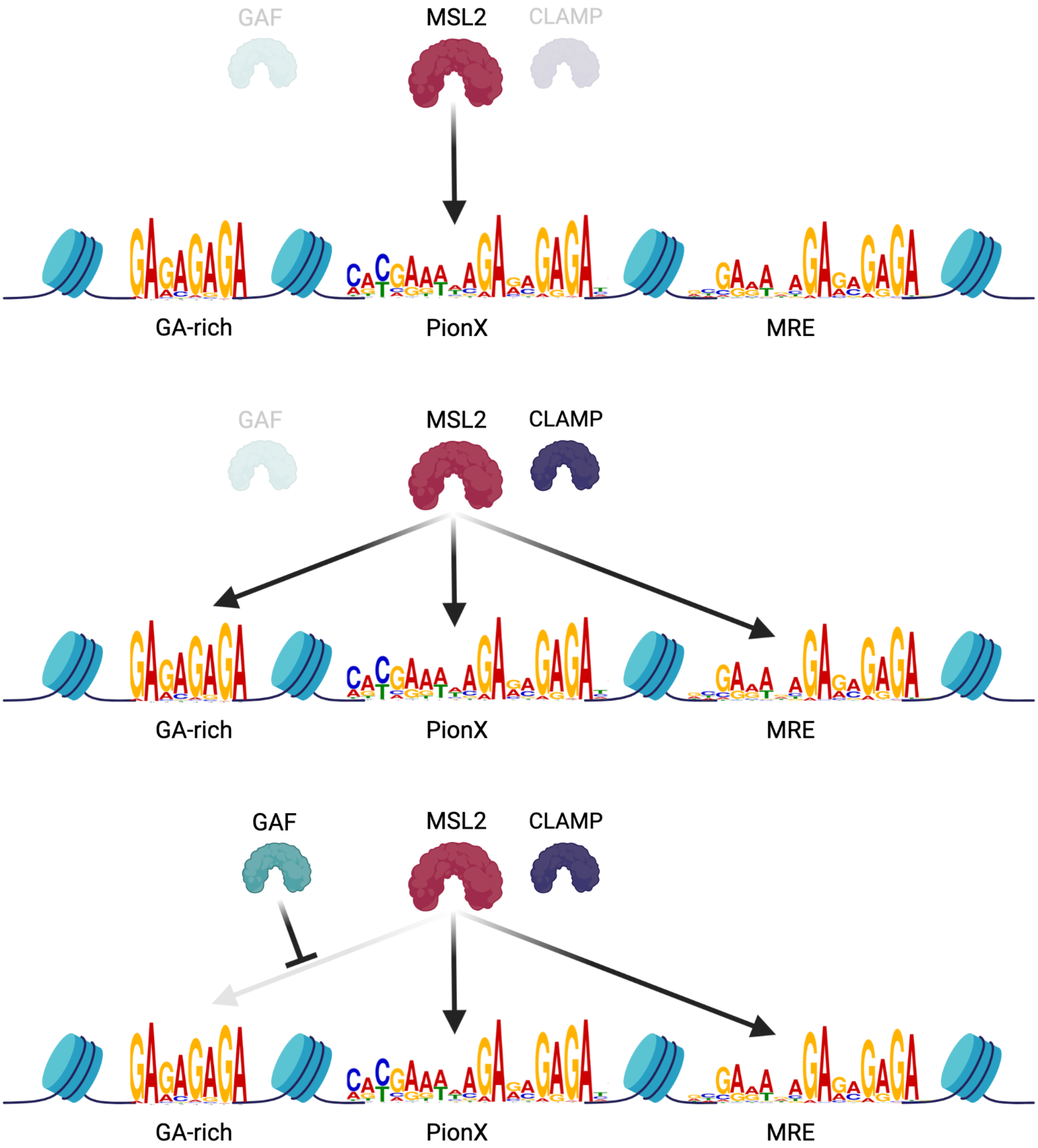


Figure 48: Model demonstrating how MSL2 binding specificity is enhanced by competition with GAF. Functional sites are not only GA-rich but harbor specific sequence and shape features. These are not recognized by MSL2 but impede GAF to bind and therefore ease competition at these sites. At nonfunctional GA-rich sites, MSL2 will bind in absence of GAF but will be outcompeted after GAF addition. Created with BioRender.com.

To analyze shape features the bound motifs from different peaks have to be aligned with base pair precision. The MRE PWM is unsuited for this kind of analysis due to its repetitive nature. Long GA stretches may be aligned sufficiently well in several ways, as it is very difficult to determine the beginning of the motif within a longer stretch. Therefore, a FIMO search using the PionX PWM was used to obtain suitable binding motifs, as the 5´extension of this motif provides an anker to which binding motifs may be aligned even if they involve GA repeats at the 3’ end. A flexible threshold was set and hits were excluded if they overlapped with a more significant hit, to assure a significant number of hits even in GAF-bound sites, while trying to avoid secondary hits within longer GA stretches. This analysis only accounts for the best hits within a peak and might be blinded to cooperative effects mediated by closeby MRE motifs within the same HAS. However, the analysis proved that MSL2-bound motifs resemble the canonical HAS shape features only in presence of GAF, which in turn shows that MSL2 itself lacks intrinsic specificity for the determined shape features. GAF, on the other hand, can only compete at sites with its minimal binding sequence GAGAG (Kaye et al., 2018; Omichinski et al., 1997). Functional sites can therefore be distinguished especially at positions 7-9, 12/13 and 17-19 from nonfunctional sites, where the GA stretches in proper HAS would end. Interestingly, these are the same locations at which sites bound by MSL2 in vitro differ significantly depending on weather they lie on the X-chromosome or autosome (Figure 38). The high variability of shapes at these locations also indicates that the specific base present is not important, provided that it is not an extension of the GA stretch. GAF improves MSL2 specificity by occluding these mostly autosomal sites through this negative selection mechanism.

4.2.4 Failures of in vitro reconstitution

Nevertheless, some problems have persisted within the in vitro reconstituted system. Regarding MSL2 some enrichment was achieved but the in vivo situation could not be reconstituted, as there were still false positive and false negative binding events.

False positive binding events are basically sites, which are bound in addition to the native ones and can be explained by a number of differences present in the in vitro system. For one, the addition of factors represents an overexpression of TFs, which might cause redistribution to lower affinity sites once all “proper” binding sites are saturated, leading to problems as described in 3.3.3. Also, roX RNA was shown to reduce binding to heterochromatin in vivo (Figueiredo et al., 2014) and could therefore help to reduce false positive signal. As MSL2 is the primary activator of roX expression and expression of roX recruits MSL2 to the X (Rattner and Meller, 2004; Valsecchi et al., 2021), inclusion of roX RNA in further experiments may lead to increased X-specificity of MSL2 IPs. Pilot experiments have so far not shown any differences in MSL2’s specificity after addition of equimolar amounts of roX. This can be explained through lack of incorporation into the DCC by the helicase MLE, which is present in the extract, or because the sites that MSL2 is suggested to bind to in absence of roX are often unmappable, such as the repetitive sequences housed in heterochromatin. Standard ChIP-seq analysis might therefore overlook these effects as MSL2 is in excess in our system and the possible binding “sink” that heterochromatin represents may be without consequence. Another possible cause for missing roX effects might be that they are only facilitated through LLPS effects, which might depend on different concentrations than those present in vitro. RoX and MSL2 can nucleate around the X chromosome in vivo (Valsecchi et al., 2021). This, again, might not happen in vitro.

Generally, many other parts of the complex are missing. This could be especially deleterious regarding MSL1. The MSL1-MSL2 complex has a different DNA binding affinity than MSL2 alone hinting at cooperative effect (Li et al., 2005a). Additionally, the DCC is thought to bind as a dimer, and dimerization is facilitated by MSL1 (Hallacli et al., 2012). As a dimer the MSL1-MSL2 complex might actually bind to HAS and then lock in place around it, which would change binding kinetics (Straub et al., 2005; Zheng et al., 2014). Future experiments should involve a reconstituted, dimeric MSL1-MSL2 complex.

MSL2 is also known to ubiquitinylate itself and other members of the complex, to maintain proper equilibrium of the components (Schunter et al., 2017; Villa et al., 2012). If and to what degree this happens in vitro is unknown, as is what effect this might have on targeting. MSL2 can also ubiquitinylate H2B in vivo, again it remains unknown whether and to which extend this occurs in the in vitro system (Wu et al., 2011).

Furthermore, it was shown that chromatin conformation can influence the protein-DNA interactions (Remus et al., 2004). Obviously DREX-assembled chromatin does not mirror the physiological folding of the chromosomes. The clustering of binding sites in vivo leads to increased local TF concentrations (Lifanov et al., 2003), which might not happen to this extent in vitro. In addition to these general findings, TAD conformations and chromosome topology influence the DCC in vivo directly (Ramirez et al., 2018; Schauer et al., 2017). Similarly, it remains unclear, how much medium-affinity binding sites in the vicinity of HAS, that are known to influence binding in vivo (Gilfillan et al., 2007), might influence the binding profile, again because the 3-dimensional makeup of the chromatin in vitro might differ from the in vivo counterpart.

4.3 Outlook

4.3.1 Additional Factors to Consider

This work demonstrates the power of a cell-free chromatin reconstitution system to probe transcription factor targeting and other chromatin interactions in vitro. It allows to manipulate the protein concentrations present more precisely and easily than in vivo systems can. However, many questions remain regarding the DCC targeting in particular and the abilities of DREX-assembled chromatin system in general. For MSL2 targeting this work demonstrates the limits inherent in incomplete reconstitutions. Faithful in vivo-like targeting was not achieved and testing of the full complex recombinantly expressed in vitro could provide further insights into the mechanisms underlying targeting. To this end, high quality protein preparations of the entire DCC have already been established in our lab (Muller et al., 2020). Pilot experiments failed to return sufficient amounts of DNA from the IP however. So far specific antibodies were used, which were raised against certain protein epitopes, which might be buried within the complex. Additional antibodies for IP will need to be tested here to obtain sufficient DNA for sequencing.

Oher hitherto unknown factors could also influence the targeting in vivo, but the in vitro system is limited to the factors that are known and can be purified. In addition to the protein components that could be added into the reconstitution assays, it would be interesting to test whether roX RNA influences the targeting, as recent research has demonstrated its effect in vivo. Expanding on previous findings using the full 5-subunit DCC plus RoX RNA might, therefore, yield interesting new results.

In vitro ChIPs also allow to easily test the effects of mutations on protein targeting. FLAG-tagged proteins can be easily mutated and purified. This allows to add them to the in vitro system without the need to stably transfect Drosophila cell lines. Therefore, also otherwise lethal mutations can be tested for their mechanistic effects.

CLAMP, for example, aids MSL2 by direct binding and nucleosome-mediated cooperativity. Using the established genomic binding maps, it would be interesting to see how MSL2 binding changes if mutations are introduced into the MSL2-CLAMP interaction domains. This could allow to tease these two interaction mechanisms apart. Additionally, little is known about the specific function of each of the zinc fingers within the zinc finger array off the CLAMP protein. Mutating or deleting one or multiple of them before addition to the in vitro assay might answer some questions about their function.

Regarding the cell-free system, many possible applications seem feasible to study genomics, from elucidating targeting mechanisms of other transcription factors to answering directly chromatin structure-related questions.

The possibility to titrate H1 into chromatin is unique to this approach and could be investigated in more detail. Studying the effects of H1 titration in MNase-seq assays for example might give insights into the precise direction of the histone H1 tails in relation to the neighboring nucleosome. To this end, precise mapping of nucleosomes and linker lengths would be necessary for multiple digestion degrees and H1 concentrations to pinpoint the miniscule changes in nucleosome occupancy distribution.

It is also interesting to see, which other cellular functions could be reconstructed in a cell-free and system. Assembly, remodeling and damage sensing abilities of the extract have been demonstrated, but other signaling pathways might also be functional.

PPPS effects were described in DREX-assembled chromatin through staining of DNA. To further the understanding of how these aggregates form it might be instructive to perform Capture-C experiments, which capture chromatin conformations at particular sites, to determine if certain chromatin regions cluster particularly often even in a cell-free environment. Similarly, fluorophore-coupled proteins could be added to the system to measure if MSL2, or any other protein of interest, is evenly distributed throughout the solution even though the DNA is not. MSL2 might cluster within the chromatin globules or, more interestingly, only within certain globules. The latter might indicate a in vivo-like clustering of binding sites in vitro. Additionally, the same approach, but targeting different proteins, could be used to test if and to which extend heterochromatin can form within the extract and whether they cluster in vitro.

Lastly, DREX-assembled chromatin could also allow for TRIM21-directed depletion through a ubiquitination pathway of targeted proteins in situ. This system was recently used in vivo to reduce protein concentrations rapidly in a dynamic system (Clift et al., 2017; Mallery et al., 2010). Depleting proteins in DREX by addition of antibodies and subsequent TRIM21-mediated degradation would allow to also test the effects of proteins already present in DREX without the need to breed mutant flies for embryo collection.

In conclusion, genome-wide chromatin reconstitution using preblastoderm embryo extracts bear further potential for interesting discoveries of chromosome structure and function.

5. References

Abraham, J., Abreu, P., Aglietta, M., Ahn, E.J., Allard, D., Allekotte, I., Allen, J., Alvarez-Muniz, J., Ambrosio, M., Anchordoqui, L.*, et al.* (2010). Measurement of the depth of maximum of extensive air showers above 10{18} eV. Phys Rev Lett *104*, 091101.

Adkins, N.L., Hagerman, T.A., and Georgel, P. (2006). GAGA protein: a multi-faceted transcription factor. Biochem Cell Biol *84*, 559-567.

Afek, A., Schipper, J.L., Horton, J., Gordan, R., and Lukatsky, D.B. (2014). Protein-DNA binding in the absence of specific base-pair recognition. Proc Natl Acad Sci U S A *111*, 17140-17145.

Akhtar, A., and Becker, P.B. (2000). Activation of transcription through histone H4 acetylation by MOF, an acetyltransferase essential for dosage compensation in Drosophila. Mol Cell *5*, 367-375.

Albig, C., Tikhonova, E., Krause, S., Maksimenko, O., Regnard, C., and Becker, P.B. (2019). Factor cooperation for chromosome discrimination in Drosophila. Nucleic Acids Research *47*, 1706–1724.

Alekseyenko, A.A., Ho, J.W., Peng, S., Gelbart, M., Tolstorukov, M.Y., Plachetka, A., Kharchenko, P.V., Jung, Y.L., Gorchakov, A.A., Larschan, E.*, et al.* (2012). Sequence-specific targeting of dosage compensation in Drosophila favors an active chromatin context. PLoS Genet *8*, e1002646.

Alekseyenko, A.A., Peng, S., Larschan, E., Gorchakov, A.A., Lee, O.K., Kharchenko, P., McGrath, S.D., Wang, C.I., Mardis, E.R., Park, P.J.*, et al.* (2008). A sequence motif within chromatin entry sites directs MSL establishment on the Drosophila X chromosome. Cell *134*, 599-609.

Alexiadis, V., Varga-Weisz, P.D., Bonte, E., Becker, P.B., and Gruss, C. (1998). In vitro chromatin remodelling by chromatin accessibility complex (CHRAC) at the SV40 origin of DNA replication. EMBO J *17*, 3428-3438.

Almouzni, G., Mechali, M., and Wolffe, A.P. (1990). Competition between transcription complex assembly and chromatin assembly on replicating DNA. EMBO J *9*, 573-582.

Almouzni, G., and Wolffe, A.P. (1993). Replication-coupled chromatin assembly is required for the repression of basal transcription in vivo. Genes Dev *7*, 2033-2047.

Ame, J.C., Spenlehauer, C., and de Murcia, G. (2004). The PARP superfamily. Bioessays *26*, 882-893.

Amrein, H., and Axel, R. (1997). Genes expressed in neurons of adult male Drosophila. Cell *88*, 459-469.

Andrews, A.J., Chen, X., Zevin, A., Stargell, L.A., and Luger, K. (2010). The histone chaperone Nap1 promotes nucleosome assembly by eliminating nonnucleosomal histone DNA interactions. Mol Cell *37*, 834-842.

Andrews, A.J., and Luger, K. (2011). Nucleosome structure(s) and stability: variations on a theme. Annu Rev Biophys *40*, 99-117.

Arendt, D., Musser, J.M., Baker, C.V.H., Bergman, A., Cepko, C., Erwin, D.H., Pavlicev, M., Schlosser, G., Widder, S., Laubichler, M.D.*, et al.* (2016). The origin and evolution of cell types. Nat Rev Genet *17*, 744-757.

Arents, G., Burlingame, R.W., Wang, B.C., Love, W.E., and Moudrianakis, E.N. (1991). The nucleosomal core histone octamer at 3.1 A resolution: a tripartite protein assembly and a left-handed superhelix. Proc Natl Acad Sci U S A *88*, 10148-10152.

Avery, O.T., Macleod, C.M., and McCarty, M. (1944). Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii. J Exp Med *79*, 137-158.

Badenhorst, P., Voas, M., Rebay, I., and Wu, C. (2002). Biological functions of the ISWI chromatin remodeling complex NURF. Genes Dev *16*, 3186-3198.

Bag, I., Dale, R.K., Palmer, C., and Lei, E.P. (2019). Correction: The zinc-finger protein CLAMP promotes gypsy chromatin insulator function in Drosophila (doi:10.1242/jcs.226092). J Cell Sci *132*.

Bai, X., Alekseyenko, A.A., and Kuroda, M.I. (2004). Sequence-specific targeting of MSL complex regulates transcription of the roX RNA genes. EMBO J *23*, 2853-2861.

Bailey, T.L., and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol *2*, 28-36.

Baldi, S., and Becker, P.B. (2013). The variant histone H2A.V of Drosophila--three roles, two guises. Chromosoma *122*, 245-258.

Baldi, S., Jain, D.S., Harpprecht, L., Zabel, A., Scheibe, M., Butter, F., Straub, T., and Becker, P.B. (2018a). Genome-wide Rules of Nucleosome Phasing in Drosophila. Mol Cell *72*, 661-672 e664.

Baldi, S., Jain, D.S., Harpprecht, L., Zabel, A., Scheibe, M., Butter, F., Straub, T., and Becker, P.B. (2018b). Genome-wide Rules of Nucleosome Phasing in Drosophila. Molecular Cell *72*, 661-672.e664.

Banani, S.F., Lee, H.O., Hyman, A.A., and Rosen, M.K. (2017). Biomolecular condensates: organizers of cellular biochemistry. Nat Rev Mol Cell Biol *18*, 285-298.

Bantignies, F., Roure, V., Comet, I., Leblanc, B., Schuettengruber, B., Bonnet, J., Tixier, V., Mas, A., and Cavalli, G. (2011). Polycomb-dependent regulatory contacts between distant Hox loci in Drosophila. Cell *144*, 214-226.

Bashaw, G.J., and Baker, B.S. (1997). The regulation of the Drosophila msl-2 gene reveals a function for Sex-lethal in translational control. Cell *89*, 789-798.

Bates, D.L., and Thomas, J.O. (1981). Histones H1 and H5: one or two molecules per nucleosome? Nucleic Acids Res *9*, 5883-5894.

Becker, P.B. (2002). Nucleosome sliding: facts and fiction. EMBO J *21*, 4749-4753.

Becker, P.B., and Horz, W. (2002). ATP-dependent nucleosome remodeling. Annu Rev Biochem *71*, 247-273.

Becker, P.B., Tsukiyama, T., and Wu, C. (1994). Chromatin assembly extracts from Drosophila embryos. Methods Cell Biol *44*, 207-223.

Becker, P.B., and Workman, J.L. (2013). Nucleosome remodeling and epigenetics. Cold Spring Harb Perspect Biol *5*.

Becker, P.B., and Wu, C. (1992). Cell-free system for assembly of transcriptionally repressed chromatin from Drosophila embryos. Mol Cell Biol *12*, 2241-2249.

Bednar, J., Garcia-Saez, I., Boopathi, R., Cutter, A.R., Papai, G., Reymer, A., Syed, S.H., Lone, I.N., Tonchev, O., Crucifix, C.*, et al.* (2017). Structure and Dynamics of a 197 bp Nucleosome in Complex with Linker Histone H1. Mol Cell *66*, 384-397 e388.

Bednar, J., Horowitz, R.A., Grigoryev, S.A., Carruthers, L.M., Hansen, J.C., Koster, A.J., and Woodcock, C.L. (1998). Nucleosomes, linker DNA, and linker histone form a unique structural motif that directs the higher-order folding and compaction of chromatin. Proc Natl Acad Sci U S A *95*, 14173-14178.

Bell, O., Schwaiger, M., Oakeley, E.J., Lienert, F., Beisel, C., Stadler, M.B., and Schubeler, D. (2010). Accessibility of the Drosophila genome discriminates PcG repression, H4K16 acetylation and replication timing. Nat Struct Mol Biol *17*, 894-900.

Belote, J.M., and Lucchesi, J.C. (1980). Male-specific lethal mutations of Drosophila melanogaster. Genetics *96*, 165-186.

Berger, S.L. (2002). Histone modifications in transcriptional regulation. Curr Opin Genet Dev *12*, 142-148.

Blank, T.A., and Becker, P.B. (1995). Electrostatic mechanism of nucleosome spacing. J Mol Biol *252*, 305-313.

Bohla, D., Herold, M., Panzer, I., Buxa, M.K., Ali, T., Demmers, J., Kruger, M., Scharfe, M., Jarek, M., Bartkuhn, M.*, et al.* (2014). A functional insulator screen identifies NURF and dREAM components to be required for enhancer-blocking. PLoS One *9*, e107765.

Bonnet, J., Lindeboom, R.G.H., Pokrovsky, D., Stricker, G., Celik, M.H., Rupp, R.A.W., Gagneur, J., Vermeulen, M., Imhof, A., and Muller, J. (2019). Quantification of Proteins and Histone Marks in Drosophila Embryos Reveals Stoichiometric Relationships Impacting Chromatin Regulation. Dev Cell *51*, 632-644 e636.

Bork, P., and Koonin, E.V. (1993). An expanding family of helicases within the 'DEAD/H' superfamily. Nucleic Acids Res *21*, 751-752.

Brownell, J.E., Zhou, J., Ranalli, T., Kobayashi, R., Edmondson, D.G., Roth, S.Y., and Allis, C.D. (1996). Tetrahymena histone acetyltransferase A: a homolog to yeast Gcn5p linking histone acetylation to gene activation. Cell *84*, 843-851.

Chetverina, D., Erokhin, M., and Schedl, P. (2021). GAGA factor: a multifunctional pioneering chromatin protein. Cell Mol Life Sci *78*, 4125-4141.

Chiu, T.P., Comoglio, F., Zhou, T., Yang, L., Paro, R., and Rohs, R. (2016). DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. Bioinformatics *32*, 1211-1213.

Chodaparambil, J.V., Barbera, A.J., Lu, X., Kaye, K.M., Hansen, J.C., and Luger, K. (2007). A charged and contoured surface on the nucleosome regulates chromatin compaction. Nat Struct Mol Biol *14*, 1105-1107.

Chow, J.C., Yen, Z., Ziesche, S.M., and Brown, C.J. (2005). Silencing of the mammalian X chromosome. Annu Rev Genomics Hum Genet *6*, 69-92.

Cirillo, L.A., Lin, F.R., Cuesta, I., Friedman, D., Jarnik, M., and Zaret, K.S. (2002). Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. Mol Cell *9*, 279-289.

Clausell, J., Happel, N., Hale, T.K., Doenecke, D., and Beato, M. (2009). Histone H1 subtypes differentially modulate chromatin condensation without preventing ATP-dependent remodeling by SWI/SNF or NURF. PLoS One *4*, e0007243.

Clift, D., McEwan, W.A., Labzin, L.I., Konieczny, V., Mogessie, B., James, L.C., and Schuh, M. (2017). A Method for the Acute and Rapid Degradation of Endogenous Proteins. Cell *171*, 1692-1706 e1618.

Climent-Canto, P., Carbonell, A., Tatarski, M., Reina, O., Bujosa, P., Font-Mateu, J., Bernues, J., Beato, M., and Azorin, F. (2020). The embryonic linker histone dBigH1 alters the functional state of active chromatin. Nucleic Acids Res *48*, 4147-4160.

Copps, K., Richman, R., Lyman, L.M., Chang, K.A., Rampersad-Ammons, J., and Kuroda, M.I. (1998). Complex formation by the Drosophila MSL proteins: role of the MSL2 RING finger in protein complex assembly. EMBO J *17*, 5409-5417.

Corona, D.F., Clapier, C.R., Becker, P.B., and Tamkun, J.W. (2002). Modulation of ISWI function by site-specific histone acetylation. EMBO Rep *3*, 242-247.

Corona, D.F., Siriaco, G., Armstrong, J.A., Snarskaya, N., McClymont, S.A., Scott, M.P., and Tamkun, J.W. (2007). ISWI regulates higher-order chromatin structure and histone H1 assembly in vivo. PLoS Biol *5*, e232.

Cosma, M.P., Tanaka, T., and Nasmyth, K. (1999). Ordered recruitment of transcription and chromatin remodeling factors to a cell cycle- and developmentally regulated promoter. Cell *97*, 299-311.

Crevel, G., and Cotterill, S. (1991). DNA replication in cell-free extracts from Drosophila melanogaster. EMBO J *10*, 4361-4369.

Croston, G.E., Kerrigan, L.A., Lira, L.M., Marshak, D.R., and Kadonaga, J.T. (1991). Sequence-specific antirepression of histone H1-mediated inhibition of basal RNA polymerase II transcription. Science *251*, 643-649.

Dahlsveen, I.K., Gilfillan, G.D., Shelest, V.I., Lamm, R., and Becker, P.B. (2006). Targeting determinants of dosage compensation in Drosophila. PLoS Genet *2*, e5.

Danilov, V.I., and Tolokh, I.S. (1984). Nature of the stacking of nucleic acid bases in water: a Monte Carlo simulation. J Biomol Struct Dyn *2*, 119-130.

De Kumar, B., and Darland, D.C. (2021). The Hox protein conundrum: The "specifics" of DNA binding for Hox proteins and their partners. Dev Biol *477*, 284-292.

Demakova, O.V., Kotlikova, I.V., Gordadze, P.R., Alekseyenko, A.A., Kuroda, M.I., and Zhimulev, I.F. (2003). The MSL complex levels are critical for its correct targeting to the chromosomes in Drosophila melanogaster. Chromosoma *112*, 103-115.

DiFiore, J.V., Ptacek, T.S., Wang, Y., Li, B., Simon, J.M., and Strahl, B.D. (2020). Unique and Shared Roles for Histone H3K36 Methylation States in Transcription Regulation Functions. Cell Rep *31*, 107751.

Dodonova, S.O., Zhu, F., Dienemann, C., Taipale, J., and Cramer, P. (2020). Nucleosome-bound SOX2 and SOX11 structures elucidate pioneer factor function. Nature *580*, 669-672.

Donovan, B.T., Chen, H., Jipa, C., Bai, L., and Poirier, M.G. (2019). Dissociation rate compensation mechanism for budding yeast pioneer transcription factors. Elife *8*.

Duan, J., Rieder, L., Colonnetta, M.M., Huang, A., McKenney, M., Watters, S., Deshpande, G., Jordan, W., Fawzi, N., and Larschan, E. (2021). CLAMP and Zelda function together to promote Drosophila zygotic genome activation. Elife *10*.

Earnshaw, W.C., and Rothfield, N. (1985). Identification of a family of human centromere proteins using autoimmune sera from patients with scleroderma. Chromosoma *91*, 313-321.

Eggers, N., and Becker, P.B. (2021). Cell-free genomics reveal intrinsic, cooperative and competitive determinants of chromatin interactions. Nucleic Acids Res.

Eltsov, M., Maclellan, K.M., Maeshima, K., Frangakis, A.S., and Dubochet, J. (2008). Analysis of cryo-electron microscopy images does not support the existence of 30-nm chromatin fibers in mitotic chromosomes in situ. Proc Natl Acad Sci U S A *105*, 19732-19737.

Erdel, F., and Rippe, K. (2018). Formation of Chromatin Subcompartments by Phase Separation. Biophys J *114*, 2262-2270.

Espinas, M.L., Jimenez-Garcia, E., Vaquero, A., Canudas, S., Bernues, J., and Azorin, F. (1999). The N-terminal POZ domain of GAGA mediates the formation of oligomers that bind DNA with high affinity and specificity. J Biol Chem *274*, 16461-16469.

Farrell, J.A., and O'Farrell, P.H. (2014). From egg to gastrula: how the cell cycle is remodeled during the Drosophila mid-blastula transition. Annu Rev Genet *48*, 269-294.

Fauth, T., Muller-Planitz, F., Konig, C., Straub, T., and Becker, P.B. (2010). The DNA binding CXC domain of MSL2 is required for faithful targeting the Dosage Compensation Complex to the X chromosome. Nucleic Acids Res *38*, 3209-3221.

Figueiredo, M.L., Kim, M., Philip, P., Allgardsson, A., Stenberg, P., and Larsson, J. (2014). Non-coding roX RNAs prevent the binding of the MSL-complex to heterochromatic regions. PLoS Genet *10*, e1004865.

Filion, G.J., van Bemmel, J.G., Braunschweig, U., Talhout, W., Kind, J., Ward, L.D., Brugman, W., de Castro, I.J., Kerkhoven, R.M., Bussemaker, H.J.*, et al.* (2010). Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. Cell *143*, 212-224.

Fuda, N.J., Guertin, M.J., Sharma, S., Danko, C.G., Martins, A.L., Siepel, A., and Lis, J.T. (2015). GAGA factor maintains nucleosome-free regions and has a role in RNA polymerase II recruitment to promoters. PLoS Genet *11*, e1005108.

Fudenberg, G., Abdennur, N., Imakaev, M., Goloborodko, A., and Mirny, L.A. (2017). Emerging Evidence of Chromosome Folding by Loop Extrusion. Cold Spring Harb Symp Quant Biol *82*, 45-55.

Fyodorov, D.V., Blower, M.D., Karpen, G.H., and Kadonaga, J.T. (2004). Acf1 confers unique activities to ACF/CHRAC and promotes the formation rather than disruption of chromatin in vivo. Genes Dev *18*, 170-183.

Fyodorov, D.V., and Kadonaga, J.T. (2003). Chromatin assembly in vitro with purified recombinant ACF and NAP-1. Methods Enzymol *371*, 499-515.

Garcia-Saez, I., Menoni, H., Boopathi, R., Shukla, M.S., Soueidan, L., Noirclerc-Savoye, M., Le Roy, A., Skoufias, D.A., Bednar, J., Hamiche, A.*, et al.* (2018). Structure of an H1-Bound 6-Nucleosome Array Reveals an Untwisted Two-Start Chromatin Fiber Conformation. Mol Cell *72*, 902-915 e907.

Gaskill, M.M., Gibson, T.J., Larson, E.D., and Harrison, M.M. (2021). GAF is essential for zygotic genome activation and chromatin accessibility in the early Drosophila embryo. Elife *10*.

Gibson, B.A., Doolittle, L.K., Schneider, M.W.G., Jensen, L.E., Gamarra, N., Henry, L., Gerlich, D.W., Redding, S., and Rosen, M.K. (2019a). Organization of Chromatin by Intrinsic and Regulated Phase Separation. Cell *179*, 470-484.e421.

Gibson, B.A., Doolittle, L.K., Schneider, M.W.G., Jensen, L.E., Gamarra, N., Henry, L., Gerlich, D.W., Redding, S., and Rosen, M.K. (2019b). Organization of Chromatin by Intrinsic and Regulated Phase Separation. Cell *179*, 470-484 e421.

Gilfillan, G.D., Konig, C., Dahlsveen, I.K., Prakoura, N., Straub, T., Lamm, R., Fauth, T., and Becker, P.B. (2007). Cumulative contributions of weak DNA determinants to targeting the Drosophila dosage compensation complex. Nucleic Acids Res *35*, 3561-3572.

Girardot, C., Scholtalbers, J., Sauer, S., Su, S.Y., and Furlong, E.E. (2016). Je, a versatile suite to handle multiplexed NGS libraries with unique molecular identifiers. BMC Bioinformatics *17*, 419.

Glikin, G.C., Ruberti, I., and Worcel, A. (1984). Chromatin assembly in Xenopus oocytes: in vitro studies. Cell *37*, 33-41.

Gordan, R., Shen, N., Dror, I., Zhou, T., Horton, J., Rohs, R., and Bulyk, M.L. (2013). Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. Cell Rep *3*, 1093-1104.

Gossett, A.J., and Lieb, J.D. (2008). DNA Immunoprecipitation (DIP) for the Determination of DNA-Binding Specificity. CSH Protoc *2008*, pdb prot4972.

Granok, H., Leibovitch, B.A., Shaffer, C.D., and Elgin, S.C. (1995). Chromatin. Ga-ga over GAGA factor. Curr Biol *5*, 238-241.

Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics *27*, 1017-1018.

Greenberg, A.J., Yanowitz, J.L., and Schedl, P. (2004). The Drosophila GAGA factor is required for dosage compensation in males and for the formation of the male-specific-lethal complex chromatin entry site at 12DE. Genetics *166*, 279-289.

Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics *32*, 2847-2849.

Guertin, M.J., Martins, A.L., Siepel, A., and Lis, J.T. (2012). Accurate prediction of inducible transcription factor binding intensities in vivo. PLoS Genet *8*, e1002610.

Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., and Noble, W.S. (2007). Quantifying similarity between motifs. Genome Biol *8*, R24.

Hallacli, E., Lipp, M., Georgiev, P., Spielman, C., Cusack, S., Akhtar, A., and Kadlec, J. (2012). Msl1-mediated dimerization of the dosage compensation complex is essential for male X-chromosome regulation in Drosophila. Mol Cell *48*, 587-600.

Hamiche, A., Sandaltzopoulos, R., Gdula, D.A., and Wu, C. (1999). ATP-dependent histone octamer sliding mediated by the chromatin remodeling complex NURF. Cell *97*, 833-842.

Hamm, D.C., and Harrison, M.M. (2018). Regulatory principles governing the maternal-to-zygotic transition: insights from Drosophila melanogaster. Open Biol *8*, 180183.

Hargreaves, D.C., and Crabtree, G.R. (2011). ATP-dependent chromatin remodeling: genetics, genomics and mechanisms. Cell Res *21*, 396-420.

Harpprecht, L. (2018). In vitro chromatin reconstitution as a tool to study H2A.V and the DNA damage response. LMU, PhD dissertation.

Harpprecht, L., Baldi, S., Schauer, T., Schmidt, A., Bange, T., Robles, M.S., Kremmer, E., Imhof, A., and Becker, P.B. (2019). A Drosophila cell-free system that senses DNA breaks and triggers phosphorylation signalling. Nucleic Acids Research.

Havas, K., Flaus, A., Phelan, M., Kingston, R., Wade, P.A., Lilley, D.M., and Owen-Hughes, T. (2000). Generation of superhelical torsion by ATP-dependent chromatin remodeling activities. Cell *103*, 1133-1142.

Hebbar, P.B., and Archer, T.K. (2007). Chromatin-dependent cooperativity between site-specific transcription factors in vivo. J Biol Chem *282*, 8284-8291.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell *38*, 576-589.

Henikoff, S., and Smith, M.M. (2015). Histone variants and epigenetics. Cold Spring Harb Perspect Biol *7*, a019364.

Henn, L., Szabo, A., Imre, L., Roman, A., Abraham, A., Vedelek, B., Nanasi, P., and Boros, I.M. (2020). Alternative linker histone permits fast paced nuclear divisions in early Drosophila embryo. Nucleic Acids Res *48*, 9007-9018.

Hug, C.B., Grimaldi, A.G., Kruse, K., and Vaquerizas, J.M. (2017). Chromatin Architecture Emerges during Zygotic Genome Activation Independent of Transcription. Cell *169*, 216-228 e219.

Hughes, A.L., Jin, Y., Rando, O.J., and Struhl, K. (2012). A functional evolutionary approach to identify determinants of nucleosome positioning: a unifying model for establishing the genome-wide pattern. Mol Cell *48*, 5-15.

Irobalieva, R.N., Fogg, J.M., Catanese, D.J., Jr., Sutthibutpong, T., Chen, M., Barker, A.K., Ludtke, S.J., Harris, S.A., Schmid, M.F., Chiu, W.*, et al.* (2015). Structural diversity of supercoiled DNA. Nat Commun *6*, 8440.

Ito, T., Bulger, M., Pazin, M.J., Kobayashi, R., and Kadonaga, J.T. (1997). ACF, an ISWI-containing and ATP-utilizing chromatin assembly and remodeling factor. Cell *90*, 145-155.

Ito, T., Levenstein, M.E., Fyodorov, D.V., Kutach, A.K., Kobayashi, R., and Kadonaga, J.T. (1999). ACF consists of two subunits, Acf1 and ISWI, that function cooperatively in the ATP-dependent catalysis of chromatin assembly. Genes Dev *13*, 1529-1539.

Iurlaro, M., Stadler, M.B., Masoni, F., Jagani, Z., Galli, G.G., and Schubeler, D. (2021). Mammalian SWI/SNF continuously restores local accessibility to chromatin. Nat Genet *53*, 279-287.

Jenuwein, T., and Allis, C.D. (2001). Translating the histone code. Science *293*, 1074-1080.

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J.M., Yan, J., Sillanpaa, M.J.*, et al.* (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Res *20*, 861-873.

Jolma, A., and Taipale, J. (2011). Methods for Analysis of Transcription Factor DNA-Binding Specificity In Vitro. Subcell Biochem *52*, 155-173.

Jolma, A., Yin, Y., Nitta, K.R., Dave, K., Popov, A., Taipale, M., Enge, M., Kivioja, T., Morgunova, E., and Taipale, J. (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature *527*, 384-388.

Jordan, W., 3rd, and Larschan, E. (2021). The zinc finger protein CLAMP promotes long-range chromatin interactions that mediate dosage compensation of the Drosophila male X-chromosome. Epigenetics Chromatin *14*, 29.

Joshi, R., Passner, J.M., Rohs, R., Jain, R., Sosinsky, A., Crickmore, M.A., Jacob, V., Aggarwal, A.K., Honig, B., and Mann, R.S. (2007). Functional specificity of a Hox protein mediated by the recognition of minor groove structure. Cell *131*, 530-543.

Judd, J., Duarte, F.M., and Lis, J.T. (2021). Pioneer-like factor GAF cooperates with PBAP (SWI/SNF) and NURF (ISWI) to regulate transcription. Genes Dev *35*, 147-156.

Kagawa, W., and Kurumizaka, H. (2021). Structural basis for DNA sequence recognition by pioneer factors in nucleosomes. Curr Opin Struct Biol *71*, 59-64.

Kamakaka, R.T., and Kadonaga, J.T. (1994). The soluble nuclear fraction, a highly efficient transcription extract from Drosophila embryos. Methods Cell Biol *44*, 225-235.

Kamakaka, R.T., Tyree, C.M., and Kadonaga, J.T. (1991). Accurate and efficient RNA polymerase II transcription with a soluble nuclear fraction derived from Drosophila embryos. Proc Natl Acad Sci U S A *88*, 1024-1028.

Kang, J.G., Hamiche, A., and Wu, C. (2002). GAL4 directs nucleosome sliding induced by NURF. EMBO J *21*, 1406-1413.

Kaplan, N., Moore, I.K., Fondufe-Mittendorf, Y., Gossett, A.J., Tillo, D., Field, Y., LeProust, E.M., Hughes, T.R., Lieb, J.D., Widom, J.*, et al.* (2009). The DNA-encoded nucleosome organization of a eukaryotic genome. Nature *458*, 362-366.

Katsani, K.R., Hajibagheri, M.A., and Verrijzer, C.P. (1999). Co-operative DNA binding by GAGA transcription factor requires the conserved BTB/POZ domain and reorganizes promoter topology. EMBO J *18*, 698-708.

Kaye, E.G., Booker, M., Kurland, J.V., Conicella, A.E., Fawzi, N.L., Bulyk, M.L., Tolstorukov, M.Y., and Larschan, E. (2018). Differential Occupancy of Two GA-Binding Proteins Promotes Targeting of the Drosophila Dosage Compensation Complex to the Male X Chromosome. Cell Rep *22*, 3227-3239.

Kelley, R.L., Meller, V.H., Gordadze, P.R., Roman, G., Davis, R.L., and Kuroda, M.I. (1999). Epigenetic spreading of the Drosophila dosage compensation complex from roX RNA genes into flanking chromatin. Cell *98*, 513-522.

Kelley, R.L., Solovyeva, I., Lyman, L.M., Richman, R., Solovyev, V., and Kuroda, M.I. (1995). Expression of msl-2 causes assembly of dosage compensation regulators on the X chromosomes and female lethality in Drosophila. Cell *81*, 867-877.

Kharchenko, P.V., Alekseyenko, A.A., Schwartz, Y.B., Minoda, A., Riddle, N.C., Ernst, J., Sabo, P.J., Larschan, E., Gorchakov, A.A., Gu, T.*, et al.* (2011). Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. Nature *471*, 480-485.

Kim, S., and Shendure, J. (2019). Mechanisms of Interplay between Transcription Factors and the 3D Genome. Mol Cell *76*, 306-319.

Kingston, R.E., and Narlikar, G.J. (1999). ATP-dependent remodeling and acetylation as regulators of chromatin fluidity. Genes Dev *13*, 2339-2352.

Korber, P., and Becker, P.B. (2010). Nucleosome dynamics and epigenetic stability. Essays Biochem *48*, 63-74.

Kornberg, R.D. (1974). Chromatin structure: a repeating unit of histones and DNA. Science *184*, 868-871.

Kornberg, R.D., and Lorch, Y. (1999). Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. Cell *98*, 285-294.

Krietenstein, N., and Rando, O.J. (2020). Mesoscale organization of the chromatin fiber. Curr Opin Genet Dev *61*, 32-36.

Krietenstein, N., Wal, M., Watanabe, S., Park, B., Peterson, C.L., Pugh, B.F., and Korber, P. (2016). Genomic Nucleosome Organization Reconstituted with Pure Proteins. Cell *167*, 709-721 e712.

Krietenstein, N., Wippo, C.J., Lieleg, C., and Korber, P. (2012). Genome-wide in vitro reconstitution of yeast chromatin with in vivo-like nucleosome positioning. Methods Enzymol *513*, 205-232.

Kuo, M.H., and Allis, C.D. (1998). Roles of histone acetyltransferases and deacetylases in gene regulation. Bioessays *20*, 615-626.

Kuroda, M.I., Hilfiker, A., and Lucchesi, J.C. (2016). Dosage Compensation in Drosophila--a Model for the Coordinate Regulation of Transcription. Genetics *204*, 435-450.

Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., and Weirauch, M.T. (2018). The Human Transcription Factors. Cell *172*, 650-665.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357-359.

Längst, G., and Becker, P.B. (2001). Nucleosome mobilization and positioning by ISWI-containing chromatin-remodeling factors. J Cell Sci *114*, 2561-2568.

Längst, G., Bonte, E.J., Corona, D.F., and Becker, P.B. (1999). Nucleosome movement by CHRAC and ISWI without disruption or trans-displacement of the histone octamer. Cell *97*, 843-852.

Laptenko, O., Beckerman, R., Freulich, E., and Prives, C. (2011). p53 binding to nucleosomes within the p21 promoter in vivo leads to nucleosome loss and transcriptional activation. Proc Natl Acad Sci U S A *108*, 10385-10390.

Larschan, E., Alekseyenko, A.A., Gortchakov, A.A., Peng, S., Li, B., Yang, P., Workman, J.L., Park, P.J., and Kuroda, M.I. (2007). MSL complex is attracted to genes marked by H3K36 trimethylation using a sequence-independent mechanism. Mol Cell *28*, 121-133.

Larschan, E., Bishop, E.P., Kharchenko, P.V., Core, L.J., Lis, J.T., Park, P.J., and Kuroda, M.I. (2011). X chromosome dosage compensation via enhanced transcriptional elongation in Drosophila. Nature *471*, 115-118.

Larschan, E., Soruco, M.M., Lee, O.K., Peng, S., Bishop, E., Chery, J., Goebel, K., Feng, J., Park, P.J., and Kuroda, M.I. (2012). Identification of chromatin-associated regulators of MSL complex targeting in Drosophila dosage compensation. PLoS Genet *8*, e1002830.

Lawrence, M., Daujat, S., and Schneider, R. (2016). Lateral Thinking: How Histone Modifications Regulate Gene Expression. Trends Genet *32*, 42-56.

Laybourn, P.J., and Kadonaga, J.T. (1991). Role of nucleosomal cores and histone H1 in regulation of transcription by RNA polymerase II. Science *254*, 238-245.

Lee, C.G., Chang, K.A., Kuroda, M.I., and Hurwitz, J. (1997). The NTPase/helicase activities of Drosophila maleless, an essential factor in dosage compensation. EMBO J *16*, 2671-2681.

Lee, H., McManus, C.J., Cho, D.Y., Eaton, M., Renda, F., Somma, M.P., Cherbas, L., May, G., Powell, S., Zhang, D.*, et al.* (2014). DNA copy number evolution in Drosophila cell lines. Genome Biol *15*, R70.

Lee, T.I., and Young, R.A. (2013). Transcriptional regulation and its misregulation in disease. Cell *152*, 1237-1251.

Leibly, D.J., Nguyen, T.N., Kao, L.T., Hewitt, S.N., Barrett, L.K., and Van Voorhis, W.C. (2012). Stabilizing additives added during cell lysis aid in the solubilization of recombinant proteins. PLoS One *7*, e52482.

Li, F., Parry, D.A., and Scott, M.J. (2005a). The amino-terminal region of Drosophila MSL1 contains basic, glycine-rich, and leucine zipper-like motifs that promote X chromosome binding, self-association, and MSL2 binding, respectively. Mol Cell Biol *25*, 8913-8924.

Li, G., Levitus, M., Bustamante, C., and Widom, J. (2005b). Rapid spontaneous accessibility of nucleosomal DNA. Nat Struct Mol Biol *12*, 46-53.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, G.P.D.P. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Li, J., Sagendorf, J.M., Chiu, T.P., Pasi, M., Perez, A., and Rohs, R. (2017). Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding. Nucleic Acids Res *45*, 12877-12887.

Li, Z., Johnson, M.R., Ke, Z., Chen, L., and Welte, M.A. (2014). Drosophila lipid droplets buffer the H2Av supply to protect early embryonic development. Curr Biol *24*, 1485-1491.

Li, Z., Thiel, K., Thul, P.J., Beller, M., Kuhnlein, R.P., and Welte, M.A. (2012). Lipid droplets control the maternal histone supply of Drosophila embryos. Curr Biol *22*, 2104-2113.

Lifanov, A.P., Makeev, V.J., Nazina, A.G., and Papatsenko, D.A. (2003). Homotypic regulatory clusters in Drosophila. Genome Res *13*, 579-588.

Lionnet, T., and Wu, C. (2021). Single-molecule tracking of transcription protein dynamics in living cells: seeing is believing, but what are we seeing? Curr Opin Genet Dev *67*, 94-102.

Liu, J., Shively, C.A., and Mitra, R.D. (2020). Quantitative analysis of transcription factor binding and expression using calling cards reporter arrays. Nucleic Acids Res *48*, e50.

Liu, X., Noll, D.M., Lieb, J.D., and Clarke, N.D. (2005). DIP-chip: rapid and accurate determination of DNA-binding specificity. Genome Res *15*, 421-427.

Lowary, P.T., and Widom, J. (1998). New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. J Mol Biol *276*, 19-42.

Lu, X., Wontakal, S.N., Emelyanov, A.V., Morcillo, P., Konev, A.Y., Fyodorov, D.V., and Skoultchi, A.I. (2009). Linker histone H1 is essential for Drosophila development, the establishment of pericentric heterochromatin, and a normal polytene chromosome structure. Genes Dev *23*, 452-465.

Lucchesi, J.C. (1978). Gene dosage compensation and the evolution of sex chromosomes. Science *202*, 711-716.

Lucchesi, J.C. (2018). Transcriptional modulation of entire chromosomes: dosage compensation. J Genet *97*, 357-364.

Lucchesi, J.C., Kelly, W.G., and Panning, B. (2005). Chromatin remodeling in dosage compensation. Annu Rev Genet *39*, 615-651.

Lucchesi, J.C., and Kuroda, M.I. (2015). Dosage Compensation inDrosophila. Cold Spring Harbor Perspectives in Biology *7*, a019398.

Luger, K., Mader, A.W., Richmond, R.K., Sargent, D.F., and Richmond, T.J. (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature *389*, 251-260.

Lusser, A., and Kadonaga, J.T. (2004). Strategies for the reconstitution of chromatin. Nat Methods *1*, 19-26.

Machida, S., Takizawa, Y., Ishimaru, M., Sugita, Y., Sekine, S., Nakayama, J.I., Wolf, M., and Kurumizaka, H. (2018). Structural Basis of Heterochromatin Formation by Human HP1. Mol Cell *69*, 385-397 e388.

Madigan, J.P., Chotkowski, H.L., and Glaser, R.L. (2002). DNA double-strand break-induced phosphorylation of Drosophila histone variant H2Av helps prevent radiation-induced apoptosis. Nucleic Acids Res *30*, 3698-3705.

Maenner, S., Muller, M., Frohlich, J., Langer, D., and Becker, P.B. (2013). ATP-dependent roX RNA remodeling by the helicase maleless enables specific association of MSL proteins. Mol Cell *51*, 174-184.

Maeshima, K., Rogge, R., Tamura, S., Joti, Y., Hikima, T., Szerlong, H., Krause, C., Herman, J., Seidel, E., DeLuca, J.*, et al.* (2016). Nucleosomal arrays self-assemble into supramolecular globular structures lacking 30-nm fibers. EMBO J *35*, 1115-1132.

Maier, V.K., Chioda, M., Rhodes, D., and Becker, P.B. (2008). ACF catalyses chromatosome movements in chromatin fibres. EMBO J *27*, 817-826.

Makowski, M.M., Gaullier, G., and Luger, K. (2020). Picking a nucleosome lock: Sequence- and structure-specific recognition of the nucleosome. J Biosci *45*.

Mallery, D.L., McEwan, W.A., Bidgood, S.R., Towers, G.J., Johnson, C.M., and James, L.C. (2010). Antibodies mediate intracellular immunity through tripartite motif-containing 21 (TRIM21). Proc Natl Acad Sci U S A *107*, 19985-19990.

Mathelier, A., Xin, B., Chiu, T.P., Yang, L., Rohs, R., and Wasserman, W.W. (2016). DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo. Cell Syst *3*, 278-286 e274.

Mavrich, T.N., Ioshikhes, I.P., Venters, B.J., Jiang, C., Tomsho, L.P., Qi, J., Schuster, S.C., Albert, I., and Pugh, B.F. (2008). A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. Genome Res *18*, 1073-1083.

McCall, M., Brown, T., and Kennard, O. (1985). The crystal structure of d(G-G-G-G-C-C-C-C). A model for poly(dG).poly(dC). J Mol Biol *183*, 385-396.

McPherson, C.E., Shim, E.Y., Friedman, D.S., and Zaret, K.S. (1993). An active tissue-specific enhancer and bound transcription factors existing in a precisely positioned nucleosomal array. Cell *75*, 387-398.

McSwiggen, D.T., Mir, M., Darzacq, X., and Tjian, R. (2019). Evaluating phase separation in live cells: diagnosis, caveats, and functional consequences. Genes Dev *33*, 1619-1634.

Meldi, L., and Brickner, J.H. (2011). Compartmentalization of the nucleus. Trends Cell Biol *21*, 701-708.

Meller, V.H. (2003). Initiation of dosage compensation in Drosophila embryos depends on expression of the roX RNAs. Mech Dev *120*, 759-767.

Meller, V.H., Gordadze, P.R., Park, Y., Chu, X., Stuckenholz, C., Kelley, R.L., and Kuroda, M.I. (2000). Ordered assembly of roX RNAs into MSL complexes on the dosage-compensated X chromosome in Drosophila. Curr Biol *10*, 136-143.

Meller, V.H., Wu, K.H., Roman, G., Kuroda, M.I., and Davis, R.L. (1997). roX1 RNA paints the X chromosome of male Drosophila and is regulated by the dosage compensation system. Cell *88*, 445-457.

Michael, A.K., and Thoma, N.H. (2021). Reading the chromatinized genome. Cell *184*, 3599-3611.

Michael, S.F., Kilfoil, V.J., Schmidt, M.H., Amann, B.T., and Berg, J.M. (1992). Metal binding and folding properties of a minimalist Cys2His2 zinc finger peptide. Proc Natl Acad Sci U S A *89*, 4796-4800.

Miller, J.A., and Widom, J. (2003). Collaborative competition mechanism for gene activation in vivo. Mol Cell Biol *23*, 1623-1632.

Mirny, L.A. (2010). Nucleosome-mediated cooperativity between transcription factors. Proc Natl Acad Sci U S A *107*, 22534-22539.

Mirny, L.A. (2011). The fractal globule as a model of chromatin architecture in the cell. Chromosome Res *19*, 37-51.

Morales, V., Straub, T., Neumann, M.F., Mengus, G., Akhtar, A., and Becker, P.B. (2004). Functional integration of the histone acetyltransferase MOF into the dosage compensation complex. EMBO J *23*, 2258-2268.

Moraru, M., and Schalch, T. (2019). Chromatin fiber structural motifs as regulatory hubs of genome function? Essays Biochem *63*, 123-132.

Morgunova, E., and Taipale, J. (2017). Structural perspective of cooperative transcription factor binding. Curr Opin Struct Biol *47*, 1-8.

Moshe, A., and Kaplan, T. (2017). Genome-wide search for Zelda-like chromatin signatures identifies GAF as a pioneer factor in early fly development. Epigenetics Chromatin *10*, 33.

Moyle-Heyrman, G., Tims, H.S., and Widom, J. (2011). Structural constraints in collaborative competition of transcription factors against the nucleosome. J Mol Biol *412*, 634-646.

Mulero, M.C., Wang, V.Y., Huxford, T., and Ghosh, G. (2019). Genome reading by the NF-kappaB transcription factors. Nucleic Acids Res *47*, 9967-9989.

Muller, M., Schauer, T., Krause, S., Villa, R., Thomae, A.W., and Becker, P.B. (2020). Two-step mechanism for selective incorporation of lncRNA into a chromatin modifier. Nucleic Acids Res *48*, 7483-7501.

Nakahashi, H., Kieffer Kwon, K.R., Resch, W., Vian, L., Dose, M., Stavreva, D., Hakim, O., Pruett, N., Nelson, S., Yamane, A.*, et al.* (2013). A genome-wide map of CTCF multivalency redefines the CTCF code. Cell Rep *3*, 1678-1689.

Narlikar, G.J. (2020). Phase-separation in chromatin organization. J Biosci *45*.

Narlikar, G.J., Fan, H.Y., and Kingston, R.E. (2002). Cooperation between complexes that regulate chromatin structure and transcription. Cell *108*, 475-487.

Nelson, H.C., Finch, J.T., Luisi, B.F., and Klug, A. (1987). The structure of an oligo(dA).oligo(dT) tract and its biological implications. Nature *330*, 221-226.

Nelson, T., Hsieh, T.S., and Brutlag, D. (1979). Extracts of Drosophila embryos mediate chromatin assembly in vitro. Proc Natl Acad Sci U S A *76*, 5510-5514.

Ner, S.S., Blank, T., Perez-Paralle, M.L., Grigliatti, T.A., Becker, P.B., and Travers, A.A. (2001). HMG-D and histone H1 interplay during chromatin assembly and early embryogenesis. J Biol Chem *276*, 37569-37576.

Ner, S.S., and Travers, A.A. (1994). HMG-D, the Drosophila melanogaster homologue of HMG 1 protein, is associated with early embryonic chromatin in the absence of histone H1. EMBO J *13*, 1817-1822.

Nyborg, J.K., and Peersen, O.B. (2004). That zincing feeling: the effects of EDTA on the behaviour of zinc-binding transcriptional regulators. Biochem J *381*, e3-4.

Oberbeckmann, E., Niebauer, V., Watanabe, S., Farnung, L., Moldt, M., Schmid, A., Cramer, P., Peterson, C.L., Eustermann, S., Hopfner, K.P.*, et al.* (2021). Ruler elements in chromatin remodelers set nucleosome array spacing and phasing. Nat Commun *12*, 3232.

Oh, H., Park, Y., and Kuroda, M.I. (2003). Local spreading of MSL complexes from roX genes on the Drosophila X chromosome. Genes Dev *17*, 1334-1339.

Ohno, M., Ando, T., Priest, D.G., Kumar, V., Yoshida, Y., and Taniguchi, Y. (2019). Sub-nucleosomal Genome Structure Reveals Distinct Nucleosome Folding Motifs. Cell *176*, 520-534 e525.

Olins, D.E., and Olins, A.L. (2003). Chromatin history: our view from the bridge. Nat Rev Mol Cell Biol *4*, 809-814.

Omichinski, J.G., Pedone, P.V., Felsenfeld, G., Gronenborn, A.M., and Clore, G.M. (1997). The solution structure of a specific GAGA factor-DNA complex reveals a modular binding mode. Nat Struct Biol *4*, 122-132.

Oohara, I., Suyama, A., and Wada, A. (1983). Reconstitution mechanism of nucleosome core particles mediated by poly(L-glutamic acid). Biochim Biophys Acta *741*, 322-332.

Paris, M., Kaplan, T., Li, X.Y., Villalta, J.E., Lott, S.E., and Eisen, M.B. (2013). Extensive divergence of transcription factor binding in Drosophila embryos with highly conserved gene expression. PLoS Genet *9*, e1003748.

Patel, A., Hashimoto, H., Zhang, X., and Cheng, X. (2016). Characterization of How DNA Modifications Affect DNA Binding by C2H2 Zinc Finger Proteins. Methods Enzymol *573*, 387-401.

Perez-Montero, S., Carbonell, A., Moran, T., Vaquero, A., and Azorin, F. (2013). The embryonic linker histone H1 variant of Drosophila, dBigH1, regulates zygotic genome activation. Dev Cell *26*, 578-590.

Pfaffle, P., and Jackson, V. (1990). Studies on rates of nucleosome formation with DNA under stress. J Biol Chem *265*, 16821-16829.

Platero, J.S., Csink, A.K., Quintanilla, A., and Henikoff, S. (1998). Changes in chromosomal localization of heterochromatin-binding proteins during the cell cycle in Drosophila. J Cell Biol *140*, 1297-1306.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.

Racki, L.R., Yang, J.G., Naber, N., Partensky, P.D., Acevedo, A., Purcell, T.J., Cooke, R., Cheng, Y., and Narlikar, G.J. (2009). The chromatin remodeller ACF acts as a dimeric motor to space nucleosomes. Nature *462*, 1016-1021.

Ramachandran, S., and Henikoff, S. (2016). Transcriptional Regulators Compete with Nucleosomes Post-replication. Cell *165*, 580-592.

Ramirez, F., Bhardwaj, V., Arrigoni, L., Lam, K.C., Gruning, B.A., Villaveces, J., Habermann, B., Akhtar, A., and Manke, T. (2018). High-resolution TADs reveal DNA sequences underlying genome organization in flies. Nat Commun *9*, 189.

Ramirez, F., Lingg, T., Toscano, S., Lam, K.C., Georgiev, P., Chung, H.R., Lajoie, B.R., de Wit, E., Zhan, Y., de Laat, W.*, et al.* (2015). High-Affinity Sites Form an Interaction Network to Facilitate Spreading of the MSL Complex across the X Chromosome in Drosophila. Mol Cell *60*, 146-162.

Rattner, B.P., and Meller, V.H. (2004). Drosophila male-specific lethal 2 protein controls sex-specific expression of the roX genes. Genetics *166*, 1825-1832.

Remus, D., Beall, E.L., and Botchan, M.R. (2004). DNA topology, not DNA sequence, is a critical determinant for Drosophila ORC-DNA binding. EMBO J *23*, 897-907.

Ricci, M.A., Manzo, C., Garcia-Parajo, M.F., Lakadamyali, M., and Cosma, M.P. (2015). Chromatin fibers are formed by heterogeneous groups of nucleosomes in vivo. Cell *160*, 1145-1158.

Rieder, L.E., Jordan, W.T., 3rd, and Larschan, E.N. (2019). Targeting of the Dosage-Compensated Male X-Chromosome during Early Drosophila Development. Cell Rep *29*, 4268-4275 e4262.

Rieder, L.E., Koreski, K.P., Boltz, K.A., Kuzu, G., Urban, J.A., Bowman, S.K., Zeidman, A., Jordan, W.T., 3rd, Tolstorukov, M.Y., Marzluff, W.F.*, et al.* (2017). Histone locus regulation by the Drosophila dosage compensation adaptor protein CLAMP. Genes Dev *31*, 1494-1508.

Rippe, K. (2007). Dynamic organization of the cell nucleus. Curr Opin Genet Dev *17*, 373-380.

Robinson, J.T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat Biotechnol *29*, 24-26.

Rohs, R., West, S.M., Sosinsky, A., Liu, P., Mann, R.S., and Honig, B. (2009). The role of DNA shape in protein-DNA recognition. Nature *461*, 1248-1253.

Rudnizky, S., Khamis, H., Malik, O., Melamed, P., and Kaplan, A. (2019). The base pair-scale diffusion of nucleosomes modulates binding of transcription factors. Proc Natl Acad Sci U S A *116*, 12161-12166.

Samata, M., and Akhtar, A. (2018). Dosage Compensation of the X Chromosome: A Complex Epigenetic Assignment Involving Chromatin Regulators and Long Noncoding RNAs. Annu Rev Biochem *87*, 323-350.

Sandaltzopoulos, R., Blank, T., and Becker, P.B. (1994). Transcriptional repression by nucleosomes but not H1 in reconstituted preblastoderm Drosophila chromatin. EMBO J *13*, 373-379.

Saxton, D.S., and Rine, J. (2020). Nucleosome Positioning Regulates the Establishment, Stability, and Inheritance of Heterochromatin in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A *117*, 27493-27501.

Schauer, T., Ghavi-Helm, Y., Sexton, T., Albig, C., Regnard, C., Cavalli, G., Furlong, E.E., and Becker, P.B. (2017). Chromosome topology guides the Drosophila Dosage Compensation Complex for target gene activation. EMBO Rep.

Schiessel, H., Widom, J., Bruinsma, R.F., and Gelbart, W.M. (2001). Polymer reptation and nucleosome repositioning. Phys Rev Lett *86*, 4414-4417.

Schulz, K.N., and Harrison, M.M. (2019). Mechanisms regulating zygotic genome activation. Nat Rev Genet *20*, 221-234.

Schunter, S., Villa, R., Flynn, V., Heidelberger, J.B., Classen, A.K., Beli, P., and Becker, P.B. (2017). Ubiquitylation of the acetyltransferase MOF in Drosophila melanogaster. PLoS One *12*, e0177408.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thastrom, A., Field, Y., Moore, I.K., Wang, J.P., and Widom, J. (2006). A genomic code for nucleosome positioning. Nature *442*, 772-778.

Sekiya, T., Muthurajan, U.M., Luger, K., Tulin, A.V., and Zaret, K.S. (2009). Nucleosome-binding affinity as a primary determinant of the nuclear mobility of the pioneer transcription factor FoxA. Genes Dev *23*, 804-809.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the Drosophila genome. Cell *148*, 458-472.

Shivaswamy, S., Bhinge, A., Zhao, Y., Jones, S., Hirst, M., and Iyer, V.R. (2008). Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. PLoS Biol *6*, e65.

Shogren-Knaak, M., Ishii, H., Sun, J.M., Pazin, M.J., Davie, J.R., and Peterson, C.L. (2006). Histone H4-K16 acetylation controls chromatin structure and protein interactions. Science *311*, 844-847.

Simon, R.H., and Felsenfeld, G. (1979). A new procedure for purifying histone pairs H2A + H2B and H3 + H4 from chromatin using hydroxylapatite. Nucleic Acids Res *6*, 689-696.

Smith, E.R., Allis, C.D., and Lucchesi, J.C. (2001). Linking global histone acetylation to the transcription enhancement of X-chromosomal genes in Drosophila males. J Biol Chem *276*, 31483-31486.

Smith, E.R., Pannuti, A., Gu, W., Steurnagel, A., Cook, R.G., Allis, C.D., and Lucchesi, J.C. (2000). The drosophila MSL complex acetylates histone H4 at lysine 16, a chromatin modification linked to dosage compensation. Mol Cell Biol *20*, 312-318.

Sobolewski, C.H., Klump, H.H., and Lindsey, G.G. (1993). A novel nucleosome assembly procedure (with a little help from pectin). FEBS Lett *318*, 27-29.

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

Sönmezer, C., Kleinendorst, R., Imanci, D., Barzaghi, G., Villacorta, L., Schubeler, D., Benes, V., Molina, N., and Krebs, A.R. (2020). Molecular Co-occupancy Identifies Transcription Factor Binding Cooperativity In Vivo. Mol Cell *81(2):255-267.e6*.

Soruco, M.M., Chery, J., Bishop, E.P., Siggers, T., Tolstorukov, M.Y., Leydon, A.R., Sugden, A.U., Goebel, K., Feng, J., Xia, P.*, et al.* (2013a). The CLAMP protein links the MSL complex to the X chromosome during Drosophila dosage compensation. Genes Dev *27*, 1551-1556.

Soruco, M.M.L., Chery, J., Bishop, E.P., Siggers, T., Tolstorukov, M.Y., Leydon, A.R., Sugden, A.U., Goebel, K., Feng, J., Xia, P.*, et al.* (2013b). The CLAMP protein links the MSL complex to the X chromosome during Drosophila dosage compensation. Genes & Development *27*, 1551-1556.

Soufi, A., Garcia, M.F., Jaroszewicz, A., Osman, N., Pellegrini, M., and Zaret, K.S. (2015). Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. Cell *161*, 555-568.

Spitz, F., and Furlong, E.E. (2012). Transcription factors: from enhancer binding to developmental control. Nat Rev Genet *13*, 613-626.

Stein, A., Whitlock, J.P., Jr., and Bina, M. (1979). Acidic polypeptides can assemble both histones and chromatin in vitro at physiological ionic strength. Proc Natl Acad Sci U S A *76*, 5000-5004.

Stephenson, R.A., Thomalla, J.M., Chen, L., Kolkhof, P., White, R.P., Beller, M., and Welte, M.A. (2021). Sequestration to lipid droplets promotes histone availability by preventing turnover of excess histones. Development.

Straub, T., and Becker, P.B. (2007). Dosage compensation: the beginning and end of generalization. Nat Rev Genet *8*, 47-57.

Straub, T., Grimaud, C., Gilfillan, G.D., Mitterweger, A., and Becker, P.B. (2008). The chromosomal high-affinity binding sites for the Drosophila dosage compensation complex. PLoS Genet *4*, e1000302.

Straub, T., Neumann, M.F., Prestel, M., Kremmer, E., Kaether, C., Haass, C., and Becker, P.B. (2005). Stable chromosomal association of MSL2 defines a dosage-compensated nuclear compartment. Chromosoma *114*, 352-364.

Straub, T., Zabel, A., Gilfillan, G.D., Feller, C., and Becker, P.B. (2013). Different chromatin interfaces of the Drosophila dosage compensation complex revealed by high-shear ChIP-seq. Genome Res *23*, 473-485.

Strutt, H., Cavalli, G., and Paro, R. (1997). Co-localization of Polycomb protein and GAGA factor on regulatory elements responsible for the maintenance of homeotic gene expression. EMBO J *16*, 3621-3632.

Suter, D.M. (2020). Transcription Factors and DNA Play Hide and Seek. Trends Cell Biol *30*, 491-500.

Talbert, P.B., and Henikoff, S. (2010). Histone variants--ancient wrap artists of the epigenome. Nat Rev Mol Cell Biol *11*, 264-275.

Tatei, K., Kimura, K., and Ohshima, Y. (1989). New methods to investigate ATP requirement for pre-mRNA splicing: inhibition by hexokinase/glucose or an ATP-binding site blocker. J Biochem *106*, 372-375.

Taube, J.H., Allton, K., Duncan, S.A., Shen, L., and Barton, M.C. (2010). Foxa1 functions as a pioneer transcription factor at transposable elements to activate Afp during differentiation of embryonic stem cells. J Biol Chem *285*, 16135-16144.

Teif, V.B., and Rippe, K. (2009). Predicting nucleosome positions on the DNA: combining intrinsic sequence preferences and remodeler activities. Nucleic Acids Res *37*, 5641-5655.

Thastrom, A., Lowary, P.T., Widlund, H.R., Cao, H., Kubista, M., and Widom, J. (1999). Sequence motifs and free energies of selected natural and non-natural nucleosome positioning DNA sequences. J Mol Biol *288*, 213-229.

Tikhonova, E., Fedotova, A., Bonchuk, A., Mogila, V., Larschan, E.N., Georgiev, P., and Maksimenko, O. (2019). The simultaneous interaction of MSL2 with CLAMP and DNA provides redundancy in the initiation of dosage compensation in Drosophila males. Development *146*, dev179663.

Tremethick, D.J. (2007). Higher-order structures of chromatin: the elusive 30 nm fiber. Cell *128*, 651-654.

Tsukiyama, T., Becker, P.B., and Wu, C. (1994). ATP-dependent nucleosome disruption at a heat-shock promoter mediated by binding of GAGA transcription factor. Nature *367*, 525-532.

Tsukiyama, T., and Wu, C. (1995). Purification and properties of an ATP-dependent nucleosome remodeling factor. Cell *83*, 1011-1020.

Tyagi, M., Imam, N., Verma, K., and Patel, A.K. (2016). Chromatin remodelers: We are the drivers!! Nucleus *7*, 388-404.

Tyler, J.K., and Kadonaga, J.T. (1999). The "dark side" of chromatin remodeling: repressive effects on transcription. Cell *99*, 443-446.

Urban, J., Kuzu, G., Bowman, S., Scruggs, B., Henriques, T., Kingston, R., Adelman, K., Tolstorukov, M., and Larschan, E. (2017a). Enhanced chromatin accessibility of the dosage compensated Drosophila male X-chromosome requires the CLAMP zinc finger protein. PLoS One *12*, e0186855.

Urban, J.A., Doherty, C.A., Jordan, W.T., 3rd, Bliss, J.E., Feng, J., Soruco, M.M., Rieder, L.E., Tsiarli, M.A., and Larschan, E.N. (2017b). The essential Drosophila CLAMP protein differentially regulates non-coding roX RNAs in male and females. Chromosome Res *25*, 101-113.

Urban, J.A., Urban, J.M., Kuzu, G., and Larschan, E.N. (2017c). The Drosophila CLAMP protein associates with diverse proteins on chromatin. PLoS One *12*, e0189772.

Valsecchi, C.I.K., Basilicata, M.F., Georgiev, P., Gaub, A., Seyfferth, J., Kulkarni, T., Panhale, A., Semplicio, G., Manjunath, V., Holz, H.*, et al.* (2021). RNA nucleation by MSL2 induces selective X chromosome compartmentalization. Nature *589*, 137-142.

van Steensel, B., Delrow, J., and Bussemaker, H.J. (2003). Genomewide analysis of Drosophila GAGA factor target genes reveals context-dependent DNA binding. Proc Natl Acad Sci U S A *100*, 2580-2585.

Varga-Weisz, P.D., Wilm, M., Bonte, E., Dumas, K., Mann, M., and Becker, P.B. (1997). Chromatin-remodelling factor CHRAC contains the ATPases ISWI and topoisomerase II. Nature *388*, 598-602.

Vasseur, P., Tonazzini, S., Ziane, R., Camasses, A., Rando, O.J., and Radman-Livaja, M. (2016). Dynamics of Nucleosome Positioning Maturation following Genomic Replication. Cell Rep *16*, 2651-2665.

Vettese-Dadey, M., Grant, P.A., Hebbes, T.R., Crane- Robinson, C., Allis, C.D., and Workman, J.L. (1996). Acetylation of histone H4 plays a primary role in enhancing transcription factor binding to nucleosomal DNA in vitro. EMBO J *15*, 2508-2518.

Villa, R., Forne, I., Muller, M., Imhof, A., Straub, T., and Becker, P.B. (2012). MSL2 combines sensor and effector functions in homeostatic control of the Drosophila dosage compensation machinery. Mol Cell *48*, 647-654.

Villa, R., Schauer, T., Smialowski, P., Straub, T., and Becker, P.B. (2016). PionX sites mark the X chromosome for dosage compensation. Nature *537*, 244-248.

Voelker-Albert, M.C., Pusch, M.C., Fedisch, A., Schilcher, P., Schmidt, A., and Imhof, A. (2016). A quantitative proteomic analysis of in vitro assembled chromatin. Mol Cell Proteomics, mcp.M115.053553.

Wall, G., Varga-Weisz, P.D., Sandaltzopoulos, R., and Becker, P.B. (1995). Chromatin remodeling by GAGA factor and heat shock factor at the hypersensitive Drosophila hsp26 promoter in vitro. EMBO J *14*, 1727-1736.

Wang, C.I., Alekseyenko, A.A., LeRoy, G., Elia, A.E., Gorchakov, A.A., Britton, L.M., Elledge, S.J., Kharchenko, P.V., Garcia, B.A., and Kuroda, M.I. (2013). Chromatin proteins captured by ChIP-mass spectrometry are linked to dosage compensation in Drosophila. Nat Struct Mol Biol *20*, 202-209.

Wang, Q., Sun, Q., Czajkowsky, D.M., and Shao, Z. (2018). Sub-kb Hi-C in D. melanogaster reveals conserved characteristics of TADs between insect and mammalian cells. Nat Commun *9*, 188.

Wasylyk, B., and Chambon, P. (1979). Transcription by eukaryotic RNA polymerases A and B of chromatin assembled in vitro. Eur J Biochem *98*, 317-327.

Watson, J.D., and Crick, F.H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature *171*, 737-738.

Wickham, H., Averick, M., Bryan , J., Chang, W., D'Agostino McGowan, L., Romain, F., Garrett, G., Hayes, A., Henry, L., Hester, J.*, et al.* (2019). Welcome to the tidyverse. Journal of Open Source Software *4*, 1686.

Wilkins, R.C., and Lis, J.T. (1998). GAGA factor binding to DNA via a single trinucleotide sequence element. Nucleic Acids Res *26*, 2672-2678.

Wippo, C.J., Israel, L., Watanabe, S., Hochheimer, A., Peterson, C.L., and Korber, P. (2011). The RSC chromatin remodelling enzyme has a unique role in directing the accurate positioning of nucleosomes. EMBO J *30*, 1277-1288.

Workman, J.L., Abmayr, S.M., Cromlish, W.A., and Roeder, R.G. (1988). Transcriptional regulation by the immediate early protein of pseudorabies virus during in vitro nucleosome assembly. Cell *55*, 211-219.

Wu, C. (1980). The 5' ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I. Nature *286*, 854-860.

Wu, L., Zee, B.M., Wang, Y., Garcia, B.A., and Dou, Y. (2011). The RING finger protein MSL2 in the MOF complex is an E3 ubiquitin ligase for H2B K34 and is involved in crosstalk with H3 K4 and K79 methylation. Mol Cell *43*, 132-144.

Xiao, H., Sandaltzopoulos, R., Wang, H.M., Hamiche, A., Ranallo, R., Lee, K.M., Fu, D., and Wu, C. (2001). Dual functions of largest NURF subunit NURF301 in nucleosome sliding and transcription factor interactions. Mol Cell *8*, 531-543.

Yan, C., Chen, H., and Bai, L. (2018). Systematic Study of Nucleosome-Displacing Factors in Budding Yeast. Mol Cell *71*, 294-305 e294.

Yan, J., Qiu, Y., Ribeiro Dos Santos, A.M., Yin, Y., Li, Y.E., Vinckier, N., Nariai, N., Benaglio, P., Raman, A., Li, X.*, et al.* (2021). Systematic analysis of binding of transcription factors to noncoding variants. Nature *591*, 147-151.

Yang, L., Orenstein, Y., Jolma, A., Yin, Y., Taipale, J., Shamir, R., and Rohs, R. (2017). Transcription factor family-specific DNA shape readout revealed by quantitative specificity models. Mol Syst Biol *13*, 910.

Zaret, K.S. (2020). Pioneer Transcription Factors Initiating Gene Network Changes. Annu Rev Genet *54*, 367-385.

Zaret, K.S., and Carroll, J.S. (2011). Pioneer transcription factors: establishing competence for gene expression. Genes Dev *25*, 2227-2241.

Zaret, K.S., Lerner, J., and Iwafuchi-Doi, M. (2016). Chromatin Scanning by Dynamic Binding of Pioneer Factors. Mol Cell *62*, 665-667.

Zenk, F., Zhan, Y., Kos, P., Loser, E., Atinbayeva, N., Schachtle, M., Tiana, G., Giorgetti, L., and Iovino, N. (2021). HP1 drives de novo 3D genome reorganization in early Drosophila embryos. Nature *593*, 289-293.

Zhang, R., Erler, J., and Langowski, J. (2017). Histone Acetylation Regulates Chromatin Accessibility: Role of H4K16 in Inter-nucleosome Interaction. Biophys J *112*, 450-459.

Zhang, Y. (2003). Transcriptional regulation by histone ubiquitination and deubiquitination. Genes Dev *17*, 2733-2740.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W.*, et al.* (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol *9*, R137.

Zhang, Y., Moqtaderi, Z., Rattner, B.P., Euskirchen, G., Snyder, M., Kadonaga, J.T., Liu, X.S., and Struhl, K. (2009). Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. Nat Struct Mol Biol *16*, 847-852.

Zhao, Y., and Garcia, B.A. (2015). Comprehensive Catalog of Currently Documented Histone Modifications. Cold Spring Harb Perspect Biol *7*, a025064.

Zheng, S., Villa, R., Wang, J., Feng, Y., Wang, J., Becker, P.B., and Ye, K. (2014). Structural basis of X chromosome DNA recognition by the MSL2 CXC domain duringDrosophiladosage compensation. Genes & Development *28*, 2652-2662.

Zhou, K., Gaullier, G., and Luger, K. (2019). Nucleosome structure and dynamics are coming of age. Nat Struct Mol Biol *26*, 3-13.

Zhu, F., Farnung, L., Kaasinen, E., Sahu, B., Yin, Y., Wei, B., Dodonova, S.O., Nitta, K.R., Morgunova, E., Taipale, M.*, et al.* (2018). The interaction landscape between transcription factors and the nucleosome. Nature *562*, 76-81.

6. Acknowledgements

First and Foremost, I would like to thank Prof. Peter Becker for giving me the opportunity to work on this project. His continued support and his scientific advice were invaluable and made this work possible at all. Your compassionate leadership and fervor for science have been an inspiration from day one.

Special thanks also extend to my Thesis Advisory Committee: Prof. Nicolas Gompel, PD Dr. Phillip Korber, Dr. Tobias Straub and Dr. Alessandro Baldi. All of you were incredibly helpful and supportive during the whole endeavor. I fondly remember the discussions that sparked during and after all project presentations.

I would especially like to thank Dr. Alessandro Baldi for taking on the task of mentoring me when I first came to the lab. Having someone to ask stupid questions from the start was a great help.

I also extend my gratitude to the technical assistants in our lab, especially Angelika Zabel, Silke Krause and Aline Campos-Sparr. You all have been such a great help in purifying proteins and making all those libraries. I surely would have gone mad without your support.

I also thank Dr. Stefan Krebs from LAFUGA at the Gene Centre for sequencing all my samples.

Also, I thank Dr. Paul Badenhorst for his kind gift of NURF-301 antibodies.

Additionally, I thank Dr. Andreas Thomae from the Bioimaging facility and Dr. Alessandro Scacchetti for their support in imaging and interpreting the PPPS effects in chromatin.

I would also like to thank Petra Vizjak for helping me with H1 purifications and Dr. Elisa Oberbeckmann for teaching me the protocols for SGD and octamer purification.

Regarding the bioinformatical analyses, I would like to thank Dr. Tobias Straub and Dr. Tamas Schauer from the Bioinformatics facility for providing me with my first scripts and additional advice.

I would also like to thank Dr. Elisabeth Schröder-Reiter as the coordinator of my grad school, the IRTG. The courses offered by the school were an instrumental help in the project and the excursions were always great fun.

Additionally, I would like to thank everyone from the Becker lab for making my time in the department so enjoyable and productive.

I would also like to thank Anna, Elisa, Janet, Nadia, Toby and Vera, hereafter referred to as “the chicks”. Your constant supply of highly caffeinated beverages was crucial in keeping me functional all these years.

And, of course, I thank my family for their faith in me and for supporting me at all times, even though they didn’t always get what it was that I was doing.

7. Appendix

7.1 Abbreviations

ACF	ATP-dependent chromatin assembly factor
ADP	adenosine diphosphate
ATP	adenosine triphosphate
bp	base pair
BTB	bric-à-brac, tramtrack and broad complex domain
CBD	CLAMP binding domain
CENPA	centromere protein A
CHD	Chromodomain Helicase DNA-binding
ChIP	chromatin immunoprecipitation
ChIP-seq	chromatin immunoprecipitation with high-throughput sequencing
CLAMP	chromatin linked adaptor for MSL proteins
CTCF	CCCTC-binding factor
CTD	C-terminal domain
CXC	CXC-rich motif domain
DCC	dosage compensation complex
DIP-seq	DNA immunoprecipitation with high-throughput sequencing
DNA	deoxyribonucleic acid
DGRC	Drosophila Genome Research Centre
DREX	Drosophila Embryo Extract
FIMO	find individual motif occurrence
GAF	transcription factor GAGA
HAS	high affinity site
HP1	heterochromatin protein 1
H3K36	histone 3 lysine 36
H4K16	histone 4 lysine 16
IP	immunoprecipitation
INO80	Inositol Requiring 80
ISWI	Imitation SWI
LLPS	liquid-liquid phase separation
lncRNA	long non-coding RNA
MEME	Multiple Em for Motif Elicitation
MGW	minor groove width
MLE	maleless
MNase	micrococcal nuclease
MOF	males-absent-on-the-first
MRE	MSL recognition element
mRNA	messenger ribonucleic acid
MSL	male-specific-lethal

NDR	nucleosome-depleted region
NFR	nucleosome-free region
NGS	next generation sequencing
PcG	Polycomb group
PionX	Pioneering-sites-on-the-X
Pol II	RNA Polymerase II
PPPS	polymer-polymer phase separation
ProT	propeller twist
PTM	post-transcriptional modification
PWM	position weight matrix
RING	Really Interesting New Gene
RNA	ribonucleic acid
RNA-seq	RNA sequencing
RNAi	RNA interference
RoX	RNA-on-the-X
RSC	Remodeling the Structure of Chromatin
SET2	Su(var)3-9, Enhancer-of-zeste and Trithorax-domain containing 2
SGD	salt gradient dyalisis
siRNA	small interfering RNA
SUMO	small ubiquitin-related modifier
Su(Hw)	suppressor of hairy wing
SWI/SNF	Switch/Sucrose-Non-Fermenting
SXL	sex lethal
TAD	topologically associating domain
TF	transcription factor
TRAX	Transcription Extract
TSS	transcription start site
TTS	transcription termination site
ZGA	zygotic genome activation
ZnF	zinc finger

7.2 Custom Code

7.2.1 JE Demultiplexing in Unix

```
#!/bin/sh

### sbatch instructions for the cluster

#SBATCH -J JE_demulti  # A single job name for the array

#SBATCH -p slim16               # Partition

#SBATCH -n 8                   # 12 cores

#SBATCH -N 1                    # one node ?required

#SBATCH -t 0-6:00                # Running time of 2 hours

#SBATCH --mem 20000               # Memory request

#SBATCH -o JE_%A_%a.out         # Standard output

#SBATCH -e JE_%A_%a.err         # Standard error

### read in files

FILEBASE=`ls *R1.fastq.gz | sed -e "s/_R1.fastq.gz//g"`

F1=${FILEBASE}_R1.fastq.gz

F2=${FILEBASE}_R3.fastq.gz

BF=${FILEBASE}_barcodes.txt

I1=${FILEBASE}_R2.fastq.gz

M=${FILEBASE}_JEstats.txt

java -jar  /opt/software/ngs/jemultiplexer/jemultiplexer_1.0.6_bundle.jar F1=$F1 F2=$F2
BF=$BF I1=$I1 M=$M FORCE=TRUE

### rename output

for f in *_[G,A,T,C]*[G,A,T,C]_1.txt.gz; do

   mv -- "$f" "${f%_[G,A,T,C]*[G,A,T,C]_1.txt.gz}_1.txt.gz"

done

for f in *_[G,A,T,C]*[G,A,T,C]_2.txt.gz; do

   mv -- "$f" "${f%_[G,A,T,C]*[G,A,T,C]_2.txt.gz}_2.txt.gz"

done

###piping into next script

bash bowtie2_run_PE.sh
```

7.2.2 Bowtie2 in Unix

```
### read in the files
FILES=($(ls -1 *_1.txt.gz))
### get size of array
NUMFASTQ=${#FILES[@]}
mkdir -p out
### now submit to SLURM
if [ $NUMFASTQ -ge 0 ]; then
        sbatch --array=1-$NUMFASTQ bowtie2_PE.sbatch
fi


### corresponding sbatch script bowtie2_PE.sbatch
#! /bin/bash
# bowtie.sbatch
#SBATCH -J bowtie_array   # A single job name for the array
#SBATCH -p slim16                  # Partition
#SBATCH -n 12                      # 12 cores
#SBATCH -N 6                       # one node ?required
#SBATCH -t 0-6:00                  # Running time of 2 hours
#SBATCH --mem 20000                # Memory request
#SBATCH -o out/bowtie_%A_%a.out          # Standard output
#SBATCH -e out/bowtie_%A_%a.err          # Standard error
# grab out filename
FILENAME=`ls *_1.txt.gz | head -n $SLURM_ARRAY_TASK_ID | tail -n 1`
FILEBASE=`echo ${FILENAME} | sed -e "s/_1.txt.gz//g"`
module load ngs/bowtie2
module load ngs/samtools
module load ngs/bedtools2
module load ngs/Homer
module load ngs/UCSCutils
```

```
BOWTIE_INDEX="/work/data/genomes/fly/Drosophila_melanogaster/UCSC/dm6/Sequence/Bowtie2Index/genome"

###small fragments

#BOWTIE_OPTS="-p 24 --end-to-end --very-sensitive --no-unal --no-mixed --no-discordant -I 10 -X 130"

#bowtie2 $BOWTIE_OPTS -x $BOWTIE_INDEX -1 ${FILEBASE}_1.txt.gz -2 ${FILEBASE}_2.txt.gz  > ${FILEBASE}_sub.sam 2> ${FILEBASE}_sub.stats

#samtools view -bS -@ 12 -q 10 ${FILEBASE}_sub.sam | samtools sort -@ 12 - | tee ${FILEBASE}_sub.bam | samtools index - ${FILEBASE}_sub.bam.bai

#makeTagDirectory ${FILEBASE}_sub.dir ${FILEBASE}_sub.bam -single

###nucleosome fragments

#BOWTIE_OPTS="-p 24 --end-to-end --very-sensitive --no-unal --no-mixed --no-discordant -I 130 -X 220"

#bowtie2 $BOWTIE_OPTS -x $BOWTIE_INDEX -1 ${FILEBASE}_1.txt.gz -2 ${FILEBASE}_2.txt.gz  > ${FILEBASE}_mono.sam 2> ${FILEBASE}_mono.stats

#samtools view -bS -@ 12 -q 2 ${FILEBASE}_mono.sam | samtools sort -@ 12 - | tee ${FILEBASE}_mono.bam | samtools index - ${FILEBASE}_mono.bam.bai

#makeTagDirectory ${FILEBASE}_mono.dir ${FILEBASE}_mono.bam -single

### all fragments

BOWTIE_OPTS="-p 24 --end-to-end --very-sensitive --no-unal --no-mixed --no-discordant -I 10 -X 220"

bowtie2 $BOWTIE_OPTS -x $BOWTIE_INDEX -1 ${FILEBASE}_1.txt.gz -2 ${FILEBASE}_2.txt.gz  > ${FILEBASE}_all.sam 2> ${FILEBASE}_all.stats

samtools view -bS -@ 12 -q 10 ${FILEBASE}_all.sam | samtools sort -@ 12 - | tee ${FILEBASE}_all.bam | samtools index - ${FILEBASE}_all.bam.bai

makeTagDirectory ${FILEBASE}_all.dir ${FILEBASE}_all.bam -single

rm ${FILEBASE}_all.sam

		makeUCSCfile ${FILEBASE}_all.dir -o ${FILEBASE}.bedgraph

		gunzip ${FILEBASE}.bedgraph.gz

		bedGraphToBigWig ${FILEBASE}.bedgraph /work/project/becbec_003/drosophila_genome/dm6.chromsizes.txt ${FILEBASE}.bw
```

7.2.3 Read Length Distribution Test in R

```
library(GenomicAlignments)
```

```
############ read in bam files to test read distribution
my_files <- list.files(path=".",pattern = "*.bam$")
my_chromosomes <- c("chr2L","chr2R","chr3L","chr3R","chrX","chrY","chr4")
parallel::mclapply(seq_along(my_files), mc.cores = 8, FUN = function(i){
  my_name <- gsub(".bam","",(paste("normalized",my_files[i], sep=".")))
  paired_bam <- readGAlignmentPairs(file=my_files[i])
  paired_bam <- keepSeqlevels(paired_bam, my_chromosomes, pruning.mode = "coarse")
  paired_ranges <- granges(paired_bam, on.discordant.seqnames="drop" )
  pdf( paste("frag_dens",my_name,"pdf", sep="."), width = 8, height = 8)
  #par(oma=c(5,5,5,5), mar=c(5,5,5,5), cex.lab=1.5, cex.axis=1.25, cex.main = 1.75)
  plot(density(width(paired_ranges), from=0, to=1000), lwd=2,
       main="my_name", xlab = "Frag Length")
  legend("topleft", legend = length(paired_ranges), bg = "white")
  dev.off()
})
```

7.2.4 Correlate Samples in R

```
rm(list = ls())
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))
library(ggplot2)
library(ggpmisc)
library(tidyverse)
library(GGally)
library(IRanges)
library(TxDb.Dmelanogaster.UCSC.dm6.ensGene)
library(ShortRead)
library(rtracklayer)

my_files <- list.files(path=".",pattern="*.bed$",full.name=T)
my_files <- my_files[grepl("f0",my_files)==F]
my_centers <- get(load("centers.mre_sample_sites.rda"))
```

```
#######
my_chromosomes <- c("chr2L","chr2R","chr3L","chr3R","chrX","chrY","chr4")
my_upper_limit <- 220
parallel::mclapply(seq_along(my_files), mc.cores = 8, FUN = function(i){
  my_name <- gsub("sample.","",gsub("../1peaks/","",gsub(".bed","",(paste("normalized",my_files[i], sep=".")))))
  paired_ranges <- import.bed(my_files[i])
  paired_ranges <- keepSeqlevels(paired_ranges, my_chromosomes, pruning.mode = "coarse")
  my_lengths <- seqlengths(keepSeqlevels(TxDb.Dmelanogaster.UCSC.dm6.ensGene, my_chromosomes, pruning.mode = "coarse"))
  seqlengths(paired_ranges) <- my_lengths

  paired_ranges_all <- paired_ranges[width(paired_ranges) < my_upper_limit]
  my_cov_all <- coverage(paired_ranges_all)
  total_all <- sum(as.numeric(unlist(my_cov_all)))
  norm_cov_all <- my_cov_all/total_all*10^9
  save(norm_cov_all,file=paste(my_name,"rda", sep="."))
})
coverageWindowsStranded <- function(centers, window.size = 2000, coverage) {
  #  cl <- makeCluster(getOption("cl.cores", 8))
  #  clusterExport(cl, list("centers","coverage","window.size") , envir=environment())
  centers <- centers[centers$chr %in% names(coverage),]
  #  result <- parSapply(cl, names(cov), function(x) {
  result <- sapply(names(coverage), function(x) {
    my.cov <- coverage[[x]]
    my.peaks <- centers[centers$chr==x,]
    mw.views <- Views(my.cov, start=my.peaks$center-ceiling(window.size/2), width=window.size+1)
    ## remove out-of bounds views
    flt <- start(mw.views)>0 & end(mw.views) < length(my.cov)
    mw.views <- mw.views[flt,]
    my.peaks <- my.peaks[flt,]
    if (length(mw.views) > 0) {
      mat <- as.matrix(mw.views)
      colnames(mat) <- seq(from=(0-ceiling(window.size/2)), to=0+ceiling(window.size/2))
      rownames(mat) <- rownames(my.peaks)
      return(mat)
    } else {
      return(NULL)
    }})
  mat <- Reduce(rbind, result)
  centers <- centers[rownames(centers) %in% rownames(mat),]
  na.omit(match(rownames(centers), rownames(mat)) )-> o
  centers <- centers[o,]
  mat[centers$strand=="-",] <- t(apply(mat[centers$strand=="-",],1,rev))
  mat
}
####### calculate m_area ########
my_files <- list.files(pattern="^normalized")
my_areas <- list.files(pattern = "^area")  ### my_means is still empty
parallel::mclapply(seq_along(my_files), mc.cores = 8, FUN = function(i){
  for(f in seq_along(my_centers)){
    my_name <- paste("area", my_files[i],sep=".")
    my_name <- gsub("normalized.","", my_name)
    my_name <- gsub(".all.rda","", my_name)
    my_name <- gsub("centers.","",my_name)
    norm.cover <- get(load(my_files[i]))
    center<- my_centers
    my_window = 200
    m_area <- coverageWindowsStranded(center,my_window,norm.cover)
    m_area_mean <- rowMeans(m_area)
    assign(my_name, m_area_mean)
    save(list = my_name, file = my_name)
    print(paste(my_name,"created"))
  }})
rm(list=ls())
########################################
my_set2 <- list.files(pattern="area*")
samples <- unique(str_split(my_set2,"_",simplify = T)[,2])
i=1
for(i in seq_along(samples)){
  my_set <- my_set2[grepl(paste0(samples[i],"$"),my_set2)==T]
  for(t in seq_along(my_set)){
    load(my_set[t])
  }
  my_titles <- gsub("area.","",my_set)
  my_df <-data.frame(mget(my_set))
  colnames(my_df) <- my_titles
  my_df
  my_rho <- cor.test(my_df[,1],my_df[,2], method="spearman")
  title <- my_rho[4]

  myplot <- ggpairs(my_df,axisLabels = "show")
  myplot <- myplot+labs(title=title)
  ggsave(paste0(samples[i],".correlation.pdf"),plot = myplot)
}
```

7.2.5 Subsample Replicates and Summarization

```
#!/bin/sh
module load ngs/samtools/1.9
module load ngs/Homer/4.9
module load ngs/UCSCutils
module load ngs/bedtools2
```

```
mkdir -p out
SAMPLES=`ls [0-9]*.bed | grep -v "merged" | grep -v "sample"| sed -e 's/[0-9][0-9]_\(.*\).bed/\1/'|sed -e 's/1//g'| sort| uniq `
echo ${SAMPLES}
for SAMPLE in ${SAMPLES}
    do
    A=`ls *_${SAMPLE}*.bed| grep -v "merged" | grep -v "sample"`
    for B in ${A}
        do
        wc -l ${B} >> ${SAMPLE}.txt
    done
    for B in ${A}
        do
        MIN=`cat ${SAMPLE}.txt | sort -n | awk '{print $1}' | head -n 1`
        bedtools sample -i ${B} -n ${MIN} > sample.${B}
        done
A=`ls sample*_${SAMPLE}*.bed`
cat ${A} > merged.${SAMPLE}.bed
makeTagDirectory merged.${SAMPLE}.dir ${A} -single -fragLength 150
done
INPUT=`ls merged*.dir -d | grep "neg"`
SAMPLES=`ls merged*.dir -d | grep -v "neg"`
INPUTBASE=`echo ${INPUT} | sed -e 's/.dir//g'`
SAMPLESBASE=`echo ${SAMPLES} | sed -e 's/.dir//g'`
for FILEBASE in ${SAMPLESBASE}
do
    sbatch --export=FILEBASE=$FILEBASE,INPUTBASE=$INPUTBASE 3homer.sbatch
done
```

7.2.6 Peak Calling by Homer

```
#!/bin/sh
```

```
#SBATCH -J findPeaksHomer   # A single job name for the array
#SBATCH -p slim16                    # Partition
#SBATCH -n 6                         # 6 cores
#SBATCH -N 1                         # one node ?required
#SBATCH -t 0-4:00                    # Running time of 2 hours
#SBATCH --mem 40000                  # Memory request
#SBATCH -o out/findPeaks_%A.out          # Standard output
#SBATCH -e out/findPeaks_%A.err          # Standard error


module load ngs/samtools/1.9
module load ngs/Homer/4.9
module load ngs/UCSCutils


#findPeaks ${FILEBASE}.dir  -i ${INPUTBASE}.dir -style histone -F 4 -L 2 -o ${FILEBASE}.h04.l02.txt
### -F is enrichment over input -L in enrichment over Local Background
findPeaks ${FILEBASE}.dir  -i ${INPUTBASE}.dir -style factor -size 150 -F 4 -L 2 -o ${FILEBASE}.f04.l02.txt
findPeaks ${FILEBASE}.dir  -i ${INPUTBASE}.dir -style factor -size 150 -F 6 -L 2 -o ${FILEBASE}.f06.l02.txt
findPeaks ${FILEBASE}.dir  -i ${INPUTBASE}.dir -style factor -size 150 -F 8 -L 2 -o ${FILEBASE}.f08.l02.txt
for TXT in *.txt
do
		TXTBASE=`echo ${TXT} | sed -e 's/.txt//g'`
		pos2bed.pl - ${TXTBASE}.txt > ${TXTBASE}.bed
done


		makeUCSCfile ${FILEBASE}.dir  -i ${INPUTBASE}.dir  -o ${FILEBASE}.bedgraph
		gunzip ${FILEBASE}.bedgraph.gz
		bedGraphToBigWig ${FILEBASE}.bedgraph /work/project/becbec_003/drosophila_genome/dm6.chromsizes.txt ${FILEBASE}.bw
```

7.2.7 Calculate Coverage at Genomic Windows in R

```
rm(list = ls())
library(LSD)
library(RColorBrewer)
library(TxDb.Dmelanogaster.UCSC.dm6.ensGene)
library(rtracklayer)
library(GenomicAlignments)
library(IRanges)
library(ShortRead)
library(rtracklayer)
library(grid)
library(ComplexHeatmap)
library(circlize)
library(dendextend)
library(genefilter)
library(ggplot2)
library(tidyverse)
library(zoo)
library(gridExtra)
library(matrixStats)

##########################################################################
coverageWindowsStranded <- function(centers, window.size = 2000, coverage) {
  centers <- centers[centers$chr %in% names(coverage),]
  result <- sapply(names(coverage), function(x) {
    my.cov <- coverage[[x]]
    my.peaks <- centers[centers$chr==x,]
    mw.views <- Views(my.cov, start=my.peaks$center-ceiling(window.size/2), width=window.size+1)
    ## remove out-of bounds views
    flt <- start(mw.views)>0 & end(mw.views) < length(my.cov)
    mw.views <- mw.views[flt,]
    my.peaks <- my.peaks[flt,]
    if (length(mw.views) > 0) {
      mat <- as.matrix(mw.views)
      colnames(mat) <- seq(from=(0-ceiling(window.size/2)), to=0+ceiling(window.size/2))
      rownames(mat) <- rownames(my.peaks)
      return(mat)
    } else {
      return(NULL)
    }
  })
  mat <- Reduce(rbind, result)
  centers <- centers[rownames(centers) %in% rownames(mat),]
  na.omit(match(rownames(centers), rownames(mat)) )-> o
  centers <- centers[o,]
  mat[centers$strand=="-",] <- t(apply(mat[centers$strand=="-",],1,rev))
  mat
}
############ read in bed files
my_files <- list.files(path="1peaks",pattern = "merge.*.bed$",full.names = T)
my_files <- my_files[grepl("f0",my_files)==F]
my_chromosomes <- c("chr2L","chr2R","chr3L","chr3R","chrX","chrY","chr4")
my_upper_limit <- 220
my_limit <- 125
parallel::mclapply(seq_along(my_files), mc.cores = 8, FUN = function(i){
  my_name <- gsub("merged","msl2_chip",gsub("1peaks/","",gsub(".bed","",(paste("normalized",my_files[i], sep=".")))))
  paired_ranges <- import.bed(my_files[i])
  paired_ranges <- keepSeqlevels(paired_ranges, my_chromosomes, pruning.mode = "coarse")
  my_lengths <- seqlengths(keepSeqlevels(TxDb.Dmelanogaster.UCSC.dm6.ensGene, my_chromosomes, pruning.mode = "coarse"))
  seqlengths(paired_ranges) <- my_lengths
  paired_ranges_all <- paired_ranges[width(paired_ranges) < my_upper_limit]
  paired_ranges_q1 <- paired_ranges[width(paired_ranges) >= 0   & width(paired_ranges) < my_limit]
  paired_ranges_q2 <- paired_ranges[width(paired_ranges) >= my_limit & width(paired_ranges) < my_upper_limit]
  saveRDS(paired_ranges_all, file =  paste(my_name,"all.unsized.rds", sep="."))
  paired_ranges_all <- resize(paired_ranges_all, 50, fix="center")
  paired_ranges_q1 <- resize(paired_ranges_q1, 50, fix="center")
  paired_ranges_q2 <- resize(paired_ranges_q2, 50, fix="center")
  my_cov_all <- coverage(paired_ranges_all)
  my_cov_q1 <- coverage(paired_ranges_q1)
  my_cov_q2 <- coverage(paired_ranges_q2)
  total_all <- sum(as.numeric(unlist(my_cov_all)))
  total_q1 <- sum(as.numeric(unlist(my_cov_q1)))
  total_q2 <- sum(as.numeric(unlist(my_cov_q2)))
  norm_cov_all <- my_cov_all/total_all*10^9
  norm_cov_q1  <- my_cov_q1/total_q1*10^9
  norm_cov_q2  <- my_cov_q2/total_q2*10^9
  assign(paste(my_name, "all", sep="."), norm_cov_all)
  assign(paste(my_name, "q1", sep="."), norm_cov_q1)
  assign(paste(my_name, "q2", sep="."), norm_cov_q2)
  export.bw(get(paste(my_name,"all", sep=".")), paste(my_name,"all.bw", sep="."))
  export.bw(get(paste(my_name,"q1", sep=".")), paste(my_name,"q1.bw", sep="."))
  export.bw(get(paste(my_name,"q2", sep=".")), paste(my_name,"q2.bw", sep="."))
  save(list= paste(my_name, "all", sep="."),file=paste(my_name,"all.rda", sep="."))
  save(list= paste(my_name, "q1", sep="."),file=paste(my_name,"q1.rda", sep="."))
  save(list= paste(my_name, "q2", sep="."),file=paste(my_name,"q2.rda", sep="."))
})
###for area of interest
my_centers_path <- list.files(path="../centers2", pattern = ".bed$",full.names = T) ### my_centers is 0
```

```
sapply(list.files(pattern = "^centers"), unlink)
#######
for(i in seq_along(my_centers_path)){
  my_name <- paste("centers",gsub("../centers2/","",gsub(".bed","",my_centers_path[i])),
sep=".")
  read.delim(my_centers_path[i],header = F,comment.char = "#") -> sites
  if(ncol(sites) == 6){
    sites <- sites[c(1,2,3,6)]
  }
  if(ncol(sites) == 3){
    sites[,4] <- c("+")
  }
  colnames(sites) <- c("chr","start","stop","strand")
  sites$centers <- round((sites$start+sites$stop)/2,digits = 0)
  rownames(sites) <- paste(sites$chr,sites$centers)
  assign(my_name, sites)
  save(list = my_name, file = paste(my_name, "rda", sep="."))
  print(paste(my_name,"created"))
}
my_centers <- list.files(pattern = "centers") ### my_centers is 0
#### calculated coverage means ######
my_window = 2000
my_means <- list.files(pattern = "^means")  ### my_means is still empty
my_files <- list.files(pattern ="normalized.*.rda")
my_conditions <- c("q1","q2","all")
parallel::mclapply(seq_along(my_conditions), mc.cores = 8, FUN = function(n){
  my_reps <- my_files[grepl(my_conditions[n],my_files)==T]

  parallel::mclapply(seq_along(my_reps), mc.cores = 8, FUN = function(i){
    for(f in seq_along(my_centers)){
      my_name <- paste("area", my_reps[i], my_centers[f],sep=".")
      my_name <- gsub("normalized.","", my_name)
      my_name <- gsub(".rda","", my_name)
      my_name <- gsub("centers.","", my_name)
      norm.cover <- get(load(my_reps[i]))
      center<- get(load(my_centers[f]))
      m_area <- coverageWindowsStranded(center,my_window,norm.cover)
      save(m_area, file = my_name)
      means_area <- colMeans(m_area, na.rm = FALSE, dims = 1)
      save(means_area, file = gsub("area","means",my_name))
    }}})
#############plot data######################################
library(zoo)
library(RColorBrewer)
library(dplyr)
library(ggplot2)
library(gridExtra)
rm(list=ls())
my_window = 2000
my_colors <- brewer.pal(9,"Set1") #alternative colour palette paired
smoothing <- 25
######
quartile <- c("q1","q2","all")
my_samples <- c("^means")
my_centers <- list.files(pattern = "^centers.*.rda$")

for(t in seq_along(quartile)){
  for(g in seq_along(my_samples)){
    my_means <- list.files(pattern= my_samples[g])
    my_means2 <- my_means[grepl((quartile[t]), my_means) == TRUE]
    rm(list=ls(pattern = "m1_"))
    for(h in seq_along(my_centers)){
```

```
   my_name <- paste0("m1_",quartile[t],".",gsub(".rda","",gsub("centers.","",my_centers[h])),"$")
   my_means1 <- my_means2[grepl(gsub("m1_","",my_name), my_means2) == T]
   assign(my_name,my_means1)
  }
  mymaps <- mget(ls(pattern = "m1_"))
  my_titles <- gsub("means.","",gsub(gsub("m1_",".",my_name),"",my_means1))
  my_mains <- c(gsub("m1_.","",gsub(quartile[t],"",paste(ls(pattern = "m1_")))))
  my_mains <- gsub("$","",my_mains,fixed=T)
  for(f in seq_along(mymaps)){
   my_areas <- mymaps[[f]]
   my_counter <- nrow(get(load(paste0("area.",gsub("means.","",my_areas[1])))))
   plotDF <- NULL
   for(i in seq_along(my_areas)){
    x <- get(load(my_areas[i]))
    x <- as.data.frame(x)
    x$position <- c(-1000:1000)
    x$sample <- my_titles[i]
    x$x <- rollmean(x$x,smoothing,fill = NA)
    plotDF <- bind_rows(plotDF,x)
   }
   myname <- paste0("ggplot",f)
   myplot <- ggplot(plotDF,aes(x=position, y = x, col = sample), ylab("occupancy"))+
    labs(y = "rel. occupancy",x="position [bp]",title = paste0(my_mains[f]," [",my_counter,"]"))+
    geom_area(aes(x=position, y = x),subset(plotDF,sample %in% c("input")),
         fill = "grey90", color= "grey90")+
    geom_line(aes(x=position, y = x),subset(plotDF,!sample %in% c("input")),stat="identity",size=0.5)+
    theme_classic()+
    coord_cartesian(
     #ylim = c(-1,4),
        ylim=c(min(plotDF$x,na.rm = T)*0.9,max(plotDF$x,na.rm = T)*1.1),
        xlim=c(min(plotDF$position),max(plotDF$position)),
        expand = F)+
      scale_color_manual(values=my_colors)
    assign(myname,myplot)
  }


  plots <- ls(pattern = "ggplot")
  l = mget(plots)


  ggsave(paste0("a.compplot.",quartile[t],"_",my_samples[g],".pdf"), marrangeGrob(grobs
= l, nrow=2, ncol=3), width=14,height=8)
 }}
################################
rm(list=ls())
my_window = 2000
my_colors <- brewer.pal(9,"Set1") #alternative colour palette paired
smoothing <- 25
#####################
myHeatmap <- function(mat, column_title, name, col, clustering_method_rows = "complete"){
 Heatmap(mat,
        col = col,
        column_title = column_title,
        name = name,
        show_column_names = FALSE,
        show_row_names = FALSE,
        cluster_columns = FALSE,
        cluster_rows = FALSE,
        clustering_method_rows = clustering_method_rows,
        gap = unit(1.5, "mm"),
        column_title_gp = gpar(fontsize = 6),
        column_title_rot = 90,
        row_title = paste("n =",length(BoverA)),
        row_title_rot = 90,
        row_title_gp = gpar(fontsize = 6),
        heatmap_legend_param = heatmap_legend_param,
        show_heatmap_legend = TRUE
 )}
bin.matrix <- function(m, bin.size) {
 bm <- c()
 for (i in 1:round((ncol(m)/bin.size))) {
  er <- (i*bin.size)
  sr <- er-(bin.size-1)
  bm <- cbind(bm,rowMeans(m[,sr:er]))
 }
 bm
}
color_function <- function(mat){
 x = as.vector(as.matrix(mat))
 colfun <- colorRamp2(breaks = seq(quantile(x, 0.2), quantile(x, 0.98), length = 3),
                      colors = heat.colors(3))
 return(colfun)
}
color_functionH <- function(mat){
 x = as.vector(as.matrix(mat))
 colfun <- colorRamp2(breaks = seq(quantile(x, 0.2), quantile(x, 0.95), length = 3),
                      colors = c("blue","white","red"))
 return(colfun)
}
heatmap_legend_param = list(legend_direction = "vertical",
                            legend_height = unit(3,"cm"),
                            title = "scale",
                  title_gp = gpar(fontsize = 6))
################################heatmap
quartile <- c("all")
my_samples <- c("^area")
my_centers <- list.files(pattern = "^centers.*.rda$")

for(t in seq_along(quartile)){
  for(g in seq_along(my_samples)){
    my_means <- list.files(pattern= my_samples[g])
    my_means2 <- my_means[grepl((quartile[t]), my_means) == TRUE]
    #my_means2 <- my_means2[grepl((my_samples[g]), my_means2) == TRUE]
    for(h in seq_along(my_centers)){
      my_name <- paste0("m1_",gsub(".rda","",gsub("centers.","",my_centers[h])))
      my_means1 <- my_means2[grepl(gsub("m1_","",my_name), my_means2) == T]
      assign(my_name,my_means1)
    }
    mymaps <- mget(ls(pattern = "m1_"))
    my_titles <- gsub("area.","",gsub(gsub("m1_",".",my_name),"",my_means1))
    my_mains <- c(gsub("m1_","",paste(ls(pattern = "m1_"))))
    for(f in seq_along(mymaps)){
      my_area <- mymaps[[f]]
      my_sorter <- my_area[2]
      rowMeans(get(load(my_sorter))[,950:1050])-> NFR # 100 bp binding
      m <- NFR[is.finite(NFR)]
      m.sort <- sort(m, decreasing = T, na.last = T)
      BoverA <- names(m.sort)
      for(i in seq_along(my_area)){
        my_name <- paste("ht",i,sep="")
        mat.rng <- get(load(my_area[i]))
        match(BoverA, rownames(mat.rng), nomatch = NA) -> sorted.number # gene sorting
        mat.rng <- na.omit(mat.rng[sorted.number,])
        mat.rng <- bin.matrix(mat.rng,10)
        my_ht <- myHeatmap(mat.rng, col =  color_function(get(load(my_area[1]))), column_title = my_titles[i])
        assign(my_name, my_ht)
        rm(my_ht)
      }
      my_list <- c(paste0("ht",seq(1,length(my_area))))
      my_list1 <- NULL
      for(q in seq_along(my_list)){
        x <- get(my_list[q])
        my_list1 <- my_list1+x
      }
      png(paste0("heatmap.",my_mains[f],quartile[t],".png"),units = "px",width =5000 ,height=5000,res=600)
      draw(my_list1,
           padding = unit(c(1, 1, 1, 1), "cm"))
      dev.off()
    }
  }
}
```

7.2.8 Plot Heatmaps and Mean in R

```
rm(list = ls())
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))
library(LSD)
library(RColorBrewer)
library(ComplexHeatmap)
library(genefilter)
library(ggplot2)
library(zoo)
library(gridExtra)
library(circlize)
library(IRanges)
library(ShortRead)
library(rtracklayer)
my_window = 2000
############# define heatmap function
myHeatmap <- function(mat, column_title, name, col1, clustering_method_rows = "complete",limit,i){
  Heatmap(mat,
          col = col1,
          column_title = column_title,
          name = name,
          use_raster = T,
          gap = 0,
          show_column_names = FALSE,
          show_row_names = FALSE,
          cluster_columns = FALSE,
          cluster_rows = FALSE,
          clustering_method_rows = clustering_method_rows,
          column_title_gp = gpar(fontsize = 12),
          column_title_rot = 0,
          row_title = paste("n =",length(BoverA)),
          row_title_rot = 90,
          row_title_gp = gpar(fontsize = 12),
          heatmap_legend_param = list(legend_width = unit(2.5,"cm"),
                                      title = "scale",
                                      direction = "horizontal",
                                      title_gp = gpar(fontsize = 12),
                                      labels_gp = gpar(fontsize = 9)),
          show_heatmap_legend = TRUE,
          width = 300,
        na_col = "black",
        top_annotation=HeatmapAnnotation(sum = anno_lines(colMeans(mat, na.rm =
FALSE, dims = 1),height = unit(50,"points"),
                                              ylim=c(min(limit),max(limit)),
                                              axis=if(i==1){T}else{F},
                                              border = F,gp = gpar(lwd=3,col = col1(max(mat)))),
                                   show_annotation_name = F),
        bottom_annotation=columnAnnotation(foo =
anno_text(x=c(rep("",ncol(mat)/200*40),"-1
kb",rep("",ncol(mat)/200*67),"0",rep("",ncol(mat)/200*85),"+ 1 kb",rep("",ncol(mat)/200*5)),
                                               rot=0,
                                               gp =gpar(fontsize = 9)))
  )}
### define colors
bin.matrix <- function(m, bin.size) {
  bm <- c()
  for (i in 1:round((ncol(m)/bin.size))) {
    er <- (i*bin.size)
    sr <- er-(bin.size-1)
    bm <- cbind(bm,rowMeans(m[,sr:er]))
  }
  bm
}
color_functionG <- function(mat){
  x = as.vector(as.matrix(mat))
  colfun <- colorRamp2(breaks = seq(1, quantile(x, 0.9), length = 5),
                       colors = brewer.pal(5,"BuPu"))
  return(colfun)
}
color_functionM <- function(mat){
  x = as.vector(as.matrix(mat))
  colfun <- colorRamp2(breaks = seq(1, quantile(x, 0.9), length = 4),
                      colors = brewer.pal(4,"YlGnBu"))
  return(colfun)
}
color_functionC <- function(mat){
  x = as.vector(as.matrix(mat))
  colfun <- colorRamp2(breaks = seq(1, quantile(x, 0.9), length = 4),
                      colors = brewer.pal(4,"YlGn"))
  return(colfun)
}
color_functionN <- function(mat){
  x = as.vector(as.matrix(mat))
  colfun <- colorRamp2(breaks = seq(1, quantile(x, 0.9), length = 8),
                      colors = brewer.pal(8,"YlOrBr"))
  return(colfun)
}
color_functionMN <- function(mat){
  x = as.vector(as.matrix(mat))
  colfun <- colorRamp2(breaks = seq(10, 50, length = 3),
                      colors = c("blue","white","red"))
  return(colfun)
}
############# define datasets
my_search<- gsub(".bed","",gsub("centers.","",list.files(path = "../../centers2",pattern=".bed")))
#my_search <- my_search[c(2,3,5,6,7,9,11,12,14)]
#my_search <- c("clamp_dip_sites","clamp_vivo_sites","Phaser_sites","has_sites","pionx_sites","GAF_vitro_GCM","SuHW_sites")
my_search <- c("clamp_dip")
#for(f in seq_along(my_search)){.  ### if more than 1 my_search
my_meansN <- list.files(path="../../",pattern= "^area",full.names = T)
my_meansN <- my_meansN[grepl(paste0("all.",my_search[f]), my_meansN) == TRUE][c(2,1)]
```

```
my_meansM <- list.files(path="../../",pattern= "^area",full.names = T)
my_meansC <- list.files(path="../../",pattern= "^area",full.names = T)
my_meansC <- my_meansC[grepl(paste0("all.",my_search[f]), my_meansC) == TRUE][]
my_mnase <- list.files(path="../../",pattern= "^area",full.names = T)
my_mnase <- my_mnase[grepl(paste0("sum.*.s50",my_search[f]), my_mnase) == TRUE][]

###### sort heatmaps by
my_sorter <- my_meansC[2]
###### name heatmaps
my_titlesN <- c("Nurf cntrl","Nurf +C")
my_titlesM <- c("Msl2 +M","Msl2 +C","Msl2 +C +G")
my_titlesG <- c("Gaf +C +M")
my_titlesC <- c("Clamp +C","Clamp +M")
my_titlesMN <- c("MNase +C","MNase cntrl")

rowMeans(get(load(my_sorter))[,950:1050])-> NFR # 100 bp binding
m <- NFR[is.finite(NFR)]
##### sort by chromosome x or auto
#m <- m[grepl("chrX",names(m))==F]
##### sort m for has
# df2 <- data.frame(chr=sapply(strsplit(names(m)," "),"[", 1),start=as.integer(sapply(strsplit(names(m)," "),"[", 2)))
# df2$stop <- df2$start+200
# df2$start <- df2$start-200
# df2 <- GRanges(df2)
# HAS <- import.bed("../../centers2/has_sites.bed")
# df2 <- overlapsAny(df2,HAS)
# m <- m[grepl("TRUE",df2)==T]
m.sort <- sort(m, decreasing = T, na.last = T)
BoverA <- names(m.sort)
#### select if only top hits
```

```
#BoverA <- BoverA[c(1:2000)]


if(exists("my_meansN")){
  mylimit=NULL
  for(i in seq_along(my_meansN)){
    maxsort <-
c(min(colMeans(get(load(my_meansN[i])))),max(colMeans(get(load(my_meansN[i])))))
    mylimit <- c(mylimit,maxsort)
  }
  for(i in seq_along(my_meansN)){
    mysort <- get(load(my_meansN[2]))
    my_name <- paste("htN",i,sep="")
    mat.rng <- get(load(my_meansN[i]))
    match(BoverA, rownames(mat.rng), nomatch = NA) -> sorted.number # gene sorting
    mat.rng <- na.omit(mat.rng[sorted.number,])
    mat.rng <- bin.matrix(mat.rng,10)
    my_ht <- myHeatmap(mat=mat.rng, col =  color_functionN(mysort), column_title =
my_titlesN[i],limit=mylimit,i=i)
    my_ht@matrix[,101] <- NA
    my_ht@matrix[,100] <- NA
    assign(my_name, my_ht)
  }}else{next}
if(exists("my_meansC")){
  mylimit=NULL
  for(i in seq_along(my_meansC)){
    maxsort <-
c(min(colMeans(get(load(my_meansC[i])))),max(colMeans(get(load(my_meansC[i])))))
    mylimit <- c(mylimit,maxsort)
  }
  for(i in seq_along(my_meansC)){
    mysort <- get(load(my_meansC[1]))
    my_name <- paste("htC",i,sep="")
    mat.rng <- get(load(my_meansC[i]))
    match(BoverA, rownames(mat.rng), nomatch = NA) -> sorted.number # gene sorting
    mat.rng <- na.omit(mat.rng[sorted.number,])
    mat.rng <- bin.matrix(mat.rng,10)
    my_ht <- myHeatmap(mat.rng, col =  color_functionC(mysort), column_title = my_titlesC[i],limit=mylimit,i=i)
    my_ht@matrix[,101] <- NA
    my_ht@matrix[,100] <- NA
    assign(my_name, my_ht)
  }}else{next}
if(exists("my_meansG")){
  mylimit=NULL
  for(i in seq_along(my_meansG)){
    maxsort <- c(min(colMeans(get(load(my_meansG[i])))),max(colMeans(get(load(my_meansG[i])))))
    mylimit <- c(mylimit,maxsort)
  }
  for(i in seq_along(my_meansG)){
    mysort <- get(load(my_meansG[1]))
    my_name <- paste("htG",i,sep="")
    mat.rng <- get(load(my_meansG[i]))
    match(BoverA, rownames(mat.rng), nomatch = NA) -> sorted.number # gene sorting
    mat.rng <- na.omit(mat.rng[sorted.number,])
    mat.rng <- bin.matrix(mat.rng,10)
    my_ht <- myHeatmap(mat.rng, col =  color_functionG(mysort), column_title = my_titlesG[i],limit=mylimit,i=i)
    my_ht@matrix[,101] <- NA
    my_ht@matrix[,100] <- NA
    assign(my_name, my_ht)
  }}else{next}
if(exists("my_meansM")){
  mylimit=NULL
  for(i in seq_along(my_meansM)){
    maxsort <- c(min(colMeans(get(load(my_meansM[i])))),max(colMeans(get(load(my_meansM[i])))))
    mylimit <- c(mylimit,maxsort)
  }
  for(i in seq_along(my_meansM)){
    mysort <- get(load(my_meansM[1]))
    my_name <- paste("htM",i,sep="")
    mat.rng <- get(load(my_meansM[i]))
    match(BoverA, rownames(mat.rng), nomatch = NA) -> sorted.number # gene sorting
    mat.rng <- na.omit(mat.rng[sorted.number,])
    mat.rng <- bin.matrix(mat.rng,10)
    my_ht <- myHeatmap(mat.rng, col =  color_functionM(mysort), column_title = my_titlesM[i],limit=mylimit,i=i)
    my_ht@matrix[,101] <- NA
    my_ht@matrix[,100] <- NA
    assign(my_name, my_ht)
  }}else{next}
if(exists("my_mnase")){
  mylimit=NULL
  for(i in seq_along(my_mnase)){
    maxsort <- c(min(colMeans(get(load(my_mnase[i])))),max(colMeans(get(load(my_mnase[i])))))
    mylimit <- c(mylimit,maxsort)
  }
  for(i in seq_along(my_mnase)){
    mysort <- get(load(my_mnase[2]))
    my_name <- paste("htMN",i,sep="")
    mat.rng <- get(load(my_mnase[i]))
    match(BoverA, rownames(mat.rng), nomatch = NA) -> sorted.number # gene sorting
    mat.rng <- na.omit(mat.rng[sorted.number,])
    mat.rng <- bin.matrix(mat.rng,10)
    my_ht <- myHeatmap(mat.rng, col = color_functionMN(mysort), column_title = my_titlesMN[i],limit=mylimit,i=i)
    my_ht@matrix[,101] <- NA
    my_ht@matrix[,100] <- NA
    assign(my_name, my_ht)
  }}else{next}
rm(my_ht)
### annotate heatmaps
#my_list <- (my_list[c(3,2,1,4,5)])
df <- as.integer(grepl("chrX",BoverA))
df[df=="1"]<-"Chr X"
df[df=="0"]<-"Autosome"
col_letters = c("Chr X" = "red","Autosome"="White")
hanno1 = Heatmap(df,
              cluster_rows = F,
              col=col_letters,
              column_title = "Chr X",
              column_title_rot = 90,
              width=50,
              gap = 0,
              show_column_names = FALSE,
              show_row_names = FALSE,
              cluster_columns = FALSE,
              show_heatmap_legend = F,
              heatmap_legend_param = list(title_gp = gpar(fontsize = 12,font=2),
                                    labels_gp = gpar(fontsize = 12)))

df2 <- data.frame(chr=sapply(strsplit(BoverA," "),"[", 1),start=as.integer(sapply(strsplit(BoverA," "),"[", 2)))
df2$stop <- df2$start+200
df2$start <- df2$start-200
df2 <- GRanges(df2)
```

```
HAS <- import.bed("../../centers2/has_sites.bed")
df2 <- overlapsAny(df2,HAS)
df2[df2==TRUE]<-"HAS"
df2[df2==FALSE]<-"other"
col_letters = c("HAS" = "Blue","other"="White")
hanno2 = Heatmap(df2,
                cluster_rows = F,
                col=col_letters,
                column_title = "HAS",
                column_title_rot = 90,
                width=50,
                gap=0,
                show_column_names = FALSE,
                show_row_names = FALSE,
                cluster_columns = FALSE,
                show_heatmap_legend = F,
                heatmap_legend_param = list(title_gp = gpar(fontsize = 12,font=2),
                                            labels_gp = gpar(fontsize = 12)))

###resorting heatmaps
my_list <- ls(pattern="ht")
#my_list <- my_list[c(1,2,5,6,3,4)]
my_list <- my_list[c(1,5,6,4,3)]
#my_list <- c("htC1","htC2","htMN1","htMN2","htN2","htN1","htM1","htM2","htM3","htG1")
my_list1=NULL
for(q in seq_along(my_list)){
  x <- get(my_list[q])
  my_list1 <- my_list1+x
}

my_list2 <- my_list1+hanno1+hanno2
```

```
pdf(paste0("heatmap.cmyk",my_search[f],".pdf"),colormodel='cmyk',height=8,width=length(my_list2)+8)

draw(my_list2,

    #padding = unit(c(1, 1, 1, 1), "cm"),

    column_title = paste(gsub("sites sites","sites",gsub("_"," ",my_search[f]),"sites")),

    column_title_gp = gpar(fontsize = 12))

dev.off()

#} ### close for loop if opened
```

7.2.9 Venn Diagrams in R

```
rm(list=ls())

setwd(dirname(rstudioapi::getActiveDocumentContext()$path))

library(rtracklayer)

library(GenomicRanges)

library(tidyverse)

library(IRanges)

library(ShortRead)

library(rtracklayer)

library(RColorBrewer)

library(Vennerable)

library(grid)

library(gridExtra)

library(matrixStats)

library(TxDb.Dmelanogaster.UCSC.dm6.ensGene)

library(devtools)

toMatch <- c("NOT","AND","only","resized","fasta","intersec")

x <- list.files(pattern=".bed")

x <- x[c(grep(paste(toMatch,collapse="|"),x,invert = T))]

a <- x[1]

b <- x[2]

c <- x[3]
```

```
#### distance max is 1 peak width (150bp)
A <- resize(import.bed(a),width = 250,fix="center")
B <- resize(import.bed(b),width = 250,fix="center")
C <- resize(import.bed(c),width = 250,fix="center")
############## venn
my_colors <- brewer.pal(7,"Set1")
my_peak_files <- c(A,B,C)
my_peak_file_names <- c(a,b,c)
my_pooled_peaks <- GenomicRanges::reduce(my_peak_files,ignore.strand=T)
my_pooled_peaks$peak_id <- paste("peak", seq_along(my_pooled_peaks), sep="_")


for(i in seq_along(my_peak_file_names)){
  my_overlaps <- !(is.na(findOverlaps(my_pooled_peaks, import.bed(my_peak_file_names[i]),type = "any",select = "arbitrary",maxgap = 150)))
  mcols(my_pooled_peaks) <- cbind(mcols(my_pooled_peaks), my_overlaps)
  colnames(mcols(my_pooled_peaks))[i+1] <- my_peak_file_names[i]
}
my_overlaps_df <- as.data.frame(mcols(my_pooled_peaks))
my_overlaps_df[,-1] <- data.matrix(my_overlaps_df[,-1])
my_sample_ids <- 2:ncol(my_overlaps_df)
my_overlaps_list <- sapply(my_sample_ids, function(x){my_overlaps_df$peak_id[my_overlaps_df[,x] == 1]})
names(my_overlaps_list) <- gsub(".bed","",c(a,b,c))
my_overlaps_venn <-  Venn(my_overlaps_list)
my_overlaps_venn_plot <- compute.Venn(my_overlaps_venn,doWeights=T)
### plot
pdf(paste("venn",gsub(".bed","",a),gsub(".bed","",b),gsub(".bed","",c),"pdf",sep="."), width = 6, height = 6)
plot(my_overlaps_venn_plot,show = list(Faces = FALSE))
dev.off()
```

7.2.10 X-Enrichment in R

```
rm(list=ls())
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))
library(dplyr)
library(tidyverse)
library(matrixStats)
library(TxDb.Dmelanogaster.UCSC.dm6.ensGene)
library(RColorBrewer)
#peaks to read in
my_peaks <- list.files(pattern="*.bed$")
# chromosomes to consider
my_chr <- c("chrX","chr2L","chr2R","chr3L","chr3R","chr4")
### determine color for chromosomes
chr.col <- function(feature) {
  return(switch(feature,
                "chr2L" = "#1F78B4",
                "chrY" = "#f5e618",
                "chr3L" = "#E31A1C",
                "chr3R" = "#FF7F00",
                "chr4" = "#B15928",
                "chrX" = "#6A3D9A",
                "chr2R" = "#33A02C"))
}
##read in data into tidyformat
my_names <- c(gsub("_all","",gsub("merged.","",(gsub(".bed","",my_peaks)))))
for(i in seq_along(my_peaks)){
  my_name <- paste("chip",i,sep="")
  my_chip <- read_tsv(my_peaks[i], col_names = c("chr","start","stop"),comment = "#")
  my_chip$sam <- my_names[i]
  assign(my_name, my_chip)
}
```

```
chiplist <- lapply(ls(pattern="^chip"), get)
tidychip<- as_tibble(bind_rows(chiplist))
random <- as.data.frame(seqlengths(keepSeqlevels(TxDb.Dmelanogaster.UCSC.dm6.ensGene, my_chr, pruning.mode = "coarse"))/1e6)
random$sam <- c("1 Genome")
random$chr <- rownames(random)
random1 <- random[,c(2,3,1)]
colnames(random1) <- c("sam","chr","hits")
tidychip %>%
  group_by(sam,chr) %>%
  filter(chr %in% my_chr) %>%
  summarise(hits=length(chr)) -> chiphits
plotdata <- as_tibble(rbind.data.frame(random1,chiphits))
plotdata$chr <- factor(plotdata$chr, levels=rev(my_chr))
cols <- sapply(my_chr, chr.col)
theme_bars1 <- theme_grey(base_size = 10, base_family = "") %+replace%
  theme(
    axis.title = element_text(size = rel(1.5)),
    axis.text = element_text(size = rel(1)),
    axis.ticks.x = element_line(colour = "black", size= rel(0.7)),
    axis.ticks.y = element_blank(),
    panel.background = element_rect(fill = "transparent", colour = NA),
    plot.background = element_rect(fill = "transparent", colour = NA),
    strip.background = element_blank(),
    strip.text =  element_text(size = rel(1)))
p <- ggplot(plotdata, aes(x="", y=hits, fill=chr)) + geom_bar(width=1, stat="identity") +
  facet_grid(~sam, scales = "free", space='free') +
  scale_x_discrete(expand = c(0, 0.5)) +
  coord_cartesian(expand = F)
p <- p + facet_wrap(~sam, scales ="free",strip.position = "left", nrow=length(my_peaks)+1, ncol=1)
p <- p + labs(title = "Peak distribution",x="",y="")
```

```
p1 <- p + scale_fill_manual(values=cols) + theme_bars1
p1 <- p1 + coord_flip()
p1 <- p1 + theme(strip.text.y.left = element_text(size=rel(1),angle=0))
pdf("1enrichment_X2.pdf",height=length(my_peaks)+1,width=5)
p1
dev.off()
```

7.2.11 Meme Motif Discovery

```
### set up array
#! /bin/bash
# convert bed to fasta
module load ngs/bedtools2/2.27.1

mkdir out
FILES=`find . -type f -name "*.bed" -not -name "*._*"`
echo ${FILES}
GENOME=`find ~/../../work/project/becbec_003/drosophila_genome -name
"*.BDGP6.dna.toplevel.fa" -size +10`
echo ${GENOME}
for FILE in $FILES
do
	bedtools getfasta -fi ${GENOME}	-bed $1 -fo $1.fasta
done
FASTA=`find . -name "*.fasta"`
FILENUMBER=`find . -name "*.fasta" | wc -l`
arr=($FASTA)
for i in `eval echo {1..$FILENUMBER}`
do
	F=`echo ${FASTA} | cut -d" " -f${i}`
	sbatch --export=f=${F} 2meme.sbatch
done
```

```
### calculate motifs with corresponding script 2meme.sbatch
#! /bin/bash
# findPeaks.sbatch
#SBATCH -J meme   # A single job name for the array
#SBATCH -p slim18                 # Partition
#SBATCH -n 1                      # 1 cores
#SBATCH -N 1                      # one node ?required
#SBATCH -t 0-2:00                 # Running time of 2 hours
#SBATCH --mem 8000                # Memory request
#SBATCH -o out/meme_%A.out        # Standard output
#SBATCH -e out/meme_%A.err        # Standard error
# grab out filename
module load meme/5.0.2
meme ${f} -oc ${f}memeout -mod zoops -dna -revcomp -nmotifs 2
meme ${f} -oc ${f}memeoutanr -mod anr -dna -revcomp -nmotifs 2
```

7.2.12 Motif Search by FIMO

```
#! /bin/bash
# convert bed to fasta
module load ngs/bedtools2/2.27.1
mkdir out
FILES=`find . -type f -name "*.bed" -not -name "*._*"`
echo ${FILES}
GENOME=`find ~/../../work/project/becbec_003/drosophila_genome -name
"*.BDGP6.dna.toplevel.fa" -size +10`
for FILE in $FILES
do
	bedtools getfasta -fi ${GENOME}	-bed $1 -fo $1.fasta
done
FASTA=`find . -name "*.fasta" -not -name "*._*"`
FILENUMBER=`find . -name "*.fasta" -not -name "*._*" | wc -l`
```

```
PWM=`find . -name "*pwm.txt" -not -name "*._*"`
for i in `eval echo {1..$FILENUMBER}`
do
        F=`echo ${FASTA} | cut -d" " -f${i}`
        sbatch --export=f=${F},pwm=${PWM} 2fimo.sbatch
done
### corresponding sbatch script
#! /bin/bash
#SBATCH -J fimo   # A single job name for the array
#SBATCH -p slim18                  # Partition
#SBATCH -n 1                       # 1 cores
#SBATCH -N 1                       # one node ?required
#SBATCH -t 0-2:00                  # Running time of 2 hours
#SBATCH --mem 8000                 # Memory request
#SBATCH -o out/meme_%A.out         # Standard output
#SBATCH -e out/meme_%A.err         # Standard error
module load meme/5.0.2
fimo --oc ${f}_fimo_out --qv-thresh --thresh 1e-1 ${pwm} ${f}
```

7.2.13 Convert FIMO Results to Bed Format

```
rm(list=ls())
library(rtracklayer)
library(tidyverse)
library(IRanges)
seq3 <- read.delim("../../../***/fimo.tsv",header = T)
seq4 <- seq3[which(seq3$motif_alt_id == "MEME-1"),]
seq4 <- data.frame(seq4$sequence_name,seq4$start,seq4$stop,seq4$strand)
seq4$seq4.sequence_name <- as.character(seq4$seq4.sequence_name)
seq5 <- (strsplit(seq4[,1], split = ":|-"))
n <- length(seq5[[1]])
seq5 <- structure(seq5, row.names = c(NA, -n), class = "data.frame")
```

```
seq6 <- cbind(t(seq5),seq4)
seq6[,2] <- as.integer(as.character(seq6[,2]))
seq6[,3] <- as.integer(as.character(seq6[,3]))
seq6[,2] <- seq6[,2]+seq6$seq4.start
seq6[,3] <- seq6[,2]+seq6$seq4.stop
seq7 <- seq6[,c(1,2,3,7)]
colnames(seq7) <- c("chr","start","stop","strand")
ranges <- GRanges(seq7)
ranges2 <-IRanges::reduce(ranges)
export.bed(ranges2,"vitrofimo.bed")
x <- read.delim("vitrofimo.bed",header = F)[,c(1,2,3,6)]
write_tsv(x,"vitrofimo.bed",quote_escape = F,col_names = F)
```

7.2.14 Shape Analysis

```
rm(list = ls())
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))
library(DNAshapeR)
library(BSgenome.Dmelanogaster.UCSC.dm6)
library(ggplot2)
library(dplyr)
library(tidyr)
library(ggpubr)
library(RColorBrewer)
my_files <- list.files(".",pattern="pwmin.*.bed$")
file_path <- file.path(".",my_files)
my_shapes <- c("MGW","HelT","ProT","Roll","EP")
for(f in seq_along(my_shapes)){
  fullplotdf=NULL
  plotmedians =NULL
for(i in seq_along(my_files)){
  gr <- read.delim(file_path[i],header = FALSE,comment.char = "#")
  gr <- cbind(gr[,c(1,2,3,6)])#
  myname <- gsub("bound","",gsub("pwmin","",gsub(".bed","",gsub("./","",file_path[i]))))
  strands <- as.character(gr$V6)
  strands2 <- gsub("+","plus",strands,fixed=T)
  strands2 <- gsub("-","+",strands2,fixed=T)
  strands2 <- gsub("plus","-",strands2,fixed=T)
  gr$V6 <- strands2
  colnames(gr) <- c("chr","start","stop","strand")
  #gr$strand <- c("+")
  GRanges(gr) -> grr
  grr <- resize(grr,width=(grr@ranges@width)+2,fix="end")
  getFasta(grr, Dmelanogaster,width = grr@ranges@width, filename = paste0("tmp",i))
  my_name <- paste0("pred",i)
  my_pred <- getShape(paste0("tmp",i))
  assign(my_name,my_pred)
  plotdf <- as.data.frame(my_pred[[f]])
  plotdf <- gather(plotdf,key = "position",value = "score")
  plotdf$position <- gsub("V","",plotdf$position)
  plotdf$position <- factor(plotdf$position,levels = c(1:27))
  plotdf$name <- myname
    plotdf %>%
    group_by(position,name) %>%
    summarise(median = median(score)) ->   plotdfmedians
  fullplotdf <- bind_rows(fullplotdf,plotdf)
  plotmedians <- bind_rows(plotmedians,plotdfmedians)
}
  fullplotdf$name <- factor(fullplotdf$name, levels = 
c("GAF","MSL2","MSL2+C","MSL2+G","MSL2+GC","has_sites","pionx56_sites"))
  plotmedians$name <- factor(plotmedians$name, levels = 
c("GAF","MSL2","MSL2+C","MSL2+G","MSL2+GC","has_sites","pionx56_sites"))
  title <- names(my_pred)[f]
  my_colors1 <- c("#810F7C",brewer.pal(6,"YlGnBu")[3:6],"Brown2","Brown3")
  p <- ggplot(fullplotdf,aes(x=position,y=score,fill=name,col=name))+
    geom_boxplot(width=0.9,lwd=0.1,outlier.size = 0.2,outlier.alpha = 0.05,alpha=0.4)+
    labs(title = title)+
    theme_classic()+
    scale_color_manual(values=my_colors1,aesthetics = c("color","fill"))+
    geom_path(data=plotmedians,mapping=aes(x=position,y=median,col=name,group=1),lineend = "round",size=1)
    #stat_compare_means(method="wilcox.test",aes(label = ..p.signif..),hide.ns = T) ## use if you want to compare differences at each position
 p
 ggsave(paste0(my_shapes[f],"newcol.6.pdf"), height = 6, width = 6)
}
system(command = "rm tmp*")
```