PhD degree in Systems Medicine (curriculum in Computational Biology)

European School of Molecular Medicine (SEMM),

University of Milan and University of Naples "Federico II"

Settore disciplinare: bio/11

Exploring changes in higher-order genome organisation during the coordinated transcriptional up-regulation in *drosophila* dosage compensation

Koustav Pal

IFOM, Milan

Matricola n. R11137

Supervisor: Dr. Francesco Ferrari

IFOM, Milan

Added Supervisor: Prof. Marco Foiani

IFOM, Milan

Anno accademico 2016-2017

TABLE OF CONTENTS

Differences in structural domains

Packages developed

List of Abbreviations

Technologies

NGS - Next Generation Sequencing

3C - Chromosome conformation capture

4C - Chromosome conformation capture-on-chip

5C - Chromosome conformation capture carbon copy

Hi-C - High-throughput chromosome conformation capture

TCC - Tethered Chromatin Conformation

CHiC - Capture Hi-C

ChIA-PET - Chromatin Interaction Analysis by Paired-End Tag Sequencing

CRISPR - Clustered Regularly Interspaced Short Palindromic Repeats

Genes, proteins and complexes

CLAMP - chromatin-linked adapter for MSL proteins

CTCF - CCCTC-binding factor

DCC - Dosage compensation complex

EPHA4 - ephrin type-A receptor 4

HoxD - Homeobox D cluster

IDH - Isocitrate Dehydrogenase

IHH - Indian Hedgehog

NPC - Nuclear pore complex

PAX3 - paired box gene 3

PDGFRA - Platelet Derived Growth Factor Receptor Alpha

SXL - Sex lethal

MSL - Male Sex Lethal

WNT6 - Wnt Family Member 6

Features

TADs - Topologically Associated Domains

CT - Chromosome Territories

Chemicals

DSG - Disuccinimidyl Glutarate

EGS - Ethylene glycol bis(succinimidyl succinate)

SDS - Sodium Dodecyl Sulfate

Algorithms and procedures

ICE - Iterative Correction and eigenvector decomposition procedure

Figures Index

1.0 Abstract

Dosage compensation (DC) is a highly plastic process responsible for altering transcriptional regulation, so as to preserve homeostasis in species with different karyotypes in the sexes. Over the past several decades this process has emerged as a robust model for understanding the relationship between transcriptional regulation and higher-order chromatin structure. In *Drosophila melanogaster* DC, the single male chromosome X undergoes an average two-fold transcriptional up-regulation for balancing the transcriptional output between sexes. Previous literature evidences proposed that a global change in chromosome structure may accompany this process.

Recent studies in other model systems suggested that chromosome X in response to dosage compensation shows a highly altered structure. Namely, in mammals it loses all genome compartmentalisation post silencing by *Xist,* and in *C. elegans* it shows altered insulation post reduction of gene expression. All of these studies were based on Hi-C. Yet, in case of *drosophila,* no such structural changes were found using Hi-C. This raises questions regarding the sensitivity of Hi-C in cases where transcription un-regulation is localized, and questions the mounting evidence in literature showing a causal link between transcriptional processes and higher-order chromatin structure.

Here I show that global conformational differences are indeed present in the male X chromosome and are detectable using Hi-C data on sex-sorted embryos alongside male and female cell lines. This task, was only made possible with the implementation of novel data analyses solutions. I show that the male X chromosome presents a more accessible structure. I identified differences in local genome compartmentalization, with several TAD boundaries disappearing or

weakening in male X chromosome. These boundaries co-localize with features related to the binding of the dosage compensation complex. The strongest correlation we observed was in relation to a dosage compensation complex co-factor CLAMP, which shows differential binding pattern between the sexes. This protein was reported to enhance chromatin accessibility. I present conclusive evidence supporting a changing global chromosome structure in response to dosage compensation.

I did not observe any differences in insulator binding. This is addition to change in insulation challenges the idea that insulation is a function of insulator binding. In the future, I would like to explore this avenue to understand how different players affecting genome functionality affect insulation as read-out from Hi-C data.

In the course of this work, Hi-C data binned at higher resolutions tended to become extremely memory intensive. With this, I identified a need to develop a data handling solution which would allow me to work more efficiently with such high-resolution Hi-C datasets. Although, such solutions have been described for python, no such solution exists for R. I aimed to create an on-disk database which circumvents the problem of loading data into memory, solves its own dependencies and plays well with existing Hi-C formats. To address these aims, I developed HiCLegos, a package built for the R statistical environment. HiCLegos, implements an on-disk HDF data structure for storing and manipulating Hi-C data. HiCLegos is deployed as a Bioconductor package. This ensures better dependency solving and higher visibility from a growing community of biology focused developers. Finally, HiCLegos provides methods for loading 2D matrices and consortium generated sparse matrix files. From a user perspective, HiCLegos offers analysis centred methods for data retrieval, such as retrieving data for genomic loci separated by a certain distance.

2.0 Introduction

Higher-order chromatin structure has been investigated for more than a century. This investigation and our understanding of it has been accelerated since the advent of next-generation sequencing technology and other high-throughput methods. Here, the advent of chromosome conformation capture (3C) is considered as a milestone for the study of higher-order chromatin structure.

The first part of this introduction covers, in brief the century preceding the invention of chromosome conformation capture and our understanding of the relationship between higher-order chromatin structure and genome function. The second part of this introduction covers the technique chromosome conformation capture and the advancement of our understanding of the relationship between higher-order chromatin structure and genome functionality. Lastly, I discuss transcriptional regulation, how dosage compensation is a valid model for studying transcriptional regulation and how chromatin structure effects or is affected by such regulatory processes.

2.1.0 Before chromosome conformation capture

Carl Rabl hypothesized in 1885 that chromosome structure remains conserved and the interphase chromosomes, have a certain degree of nuclear localization with an orientation that matches the polarization observed during metaphase (Cremer and Cremer 2006; 2010). These regions of localisation in the nucleus were later termed Chromosome Territories (CTs) (Cremer and Cremer 2010). Technological advancements in microscopy during the 1970s, allowed the visualisation of interphase chromosomes by Stack et al., 1977. Showcasing for the first time, the nuclear space partitioning property of interphase chromosomes(Stack et al. 1977). During this time, the popular view was that chromatin pervaded the

entirety of the nuclear space, with only heterochromatin being condensed and the euchromatic chromatin fibres being aggregated in this space (Comings 1968; Vogel and Schroeder 1974; Wischnitzer 1973). Stack et al. reconciled their findings with this model by suggesting that de-condensed chromosomes by virtue of availability of space within the nucleus will experience a certain level of cross-talk with other chromosomes at their boundaries and would result in the blurring of these boundaries. This would have been beyond or at the detectable limit of then current microscopy technologies. A scenario such as this posed the question as to the mechanistic source of chromosomal territory formation. Stack et al. postulated that a relationship between the nuclear matrix (Berezney and Coffey 1977) and de-condensed chromatin may be responsible. These results were further confirmed using different experimental procedures (Cremer et al. 1982; Zorn et al. 1979) in chinese hamster ovary cells.

Several models were laid out during the next two decades. Taking current literature into context, it was postulated that a higher-order eukaryotic genome is partitioned into three-dimensional (3D) structures characterised by a distinct differentiation state and these structures hierarchically aggregate to form a large three dimensional structure of the zygotic genome (Blobel 1985). Towards this hypothesis, certain key assumptions were made. The first being, that a 11nm “beads on a string” chromatin fiber, wound up to form a 30nm chromatin fibre, which was further packaged into higher-order hierarchically stacking structures of more or less condensed chromatin (Blobel 1985). In this model, each higher order structure was characterised by a specific differentiation state. Although very similar to what is currently known, this model had a few caveats. Since differentiation states change while transitioning through the cell cycle, these states had to converge on one common state to form the highly condensed metaphase chromosomes. To

circumvent this problem, it was proposed that these structures still existed were sub-microscopic and beyond the achievable resolution of then current technology. To ensure functional relevance of this model, it was also proposed that genotypic differences between individuals lent variation in these 3D structures, and each individual in species with dimorphic sexes had a different set of 3D structures in their germ line genomes. These 3D structure converged to form a unique zygotic ensemble from which new variant 3D structures would arise. The author theorized, that DNA by itself did not contain the complexity required to generate these three-dimensional structures. Therefore, the nuclear pore complex (NPC) and nuclear lamina were proposed as factors required for the maintenance and establishment of these three-dimensional chromatin aggregates. NPC, an organelle which acts as a bridge for macromolecular traffic between the nucleus and the cytoplasm (Feldherr et al. 1984), would act as an anchor points which hooks on onto transcribed genes in the less condensed 3D structures via DNA binding regions in its constituent subunits. Whereas, Lamins A, B, C, which make up the nuclear lamina (Gerace et al. 1978) would be responsible for structuring the high compaction regions of the chromatin. Partly in agreement with aforementioned model, it was observed that DNAase I-sensitive regions of active chromatin localized at the periphery of the interphase nucleus in cultured cells and at the the inter-chromatin space in mature red blood cell nuclei (Hutchison and Weintraub 1985).

With advances in technology, such as the development of high-resolution *in-situ* hybridization, it became possible to observe genes (Lawrence et al. 1988; Lichter et al. 1988b), chromosomal domains (Manuelidis and Borden 1988; Pinkel et al. 1986; Cremer et al. 1986), and single chromosomes (Manuelidis 1985; Pinkel et al. 1988; Lichter et al. 1988a; Manuelidis 1990). Domains could be classified into different structures based on their size and genetic constituents. In one version

(Manuelidis 1990), each genetic unit became a loop domain of approximately 30kb in size, these loop domains aggregated to form larger transcriptional and replication units that correspond to chromosomal banding patterns. In this model, even larger domains were comprised by constitutive heterochromatin regions which spanned approximately 9mb. This particular model partially aligned with the previous stated model (Blobel 1985) in the context that it took into account the hierarchical folding of chromatin and that chromatin folded into 30nm fibers. Although, the previous model did not take into account the formation of loop domains or the existence of what was proposed as the solenoid fibres. Furthermore, these small band domains on chromosomal arms could be classified based on their trypsin resistance. Trypsin resistant regions were called G-dark bands, whereas trypsin susceptible regions were called G-light bands (Holmquist 1989). G-light regions corresponded to accessible early replicating regions in lymphocytes and may host housekeeping genes (Manuelidis 1990), whilst G-dark regions corresponded to inaccessible late-replicating regions. It was noted that since non-coding DNA constitutes 90% of the genome, it may confer recognition features by creating structural partitions between functional genetic units (Manuelidis 1990). This partitioning would allow trans-acting DNA modifying and binding factors to easily reach their effector destinations (Manuelidis 1990). Already, there were also some observable functional relevance to this model. The G-dark β-globin locus is selectively turned into an accessible region in selected cell types (Dhar et al. 1989). But this locus being inappropriately transcribed in chicken brain nuclei, still showed proper β-globin expression under the control of *trans*-acting regulators (Lois et al. 1990). The G-light regions were also implicated as regions containing oncogenes (Manuelidis 1990).

Similar band domain definitions were also reached based on GC-content (Bernardi 1995). GC-rich bands were named R-bands, whereas GC-poor bands

came to be known as G-bands (Saccone et al. 1993). G and R bands have high and low gene concentrations respectively (Cuny et al. 1981; Saccone et al. 1996). Constitutively expressed housekeeping genes reside on early replicating R-bands, while late replicating G-bands contain tissue specific genes (Sadoni et al. 1999). Furthermore, it was reported that these band domains are present as distinct domains within chromosome territories (Zink et al. 1999). Additional data indicated a possible relationship between these domains and replication foci, regions where actively replicated DNA, nascent DNA and associated factors are found. By the dawn of the 21st century, the community had started to adopt an integrative view of the higher-order chromatin structure and nuclear processes such as replication and transcription (Sadoni et al. 1999).

2.2.0 Chromosome conformation capture

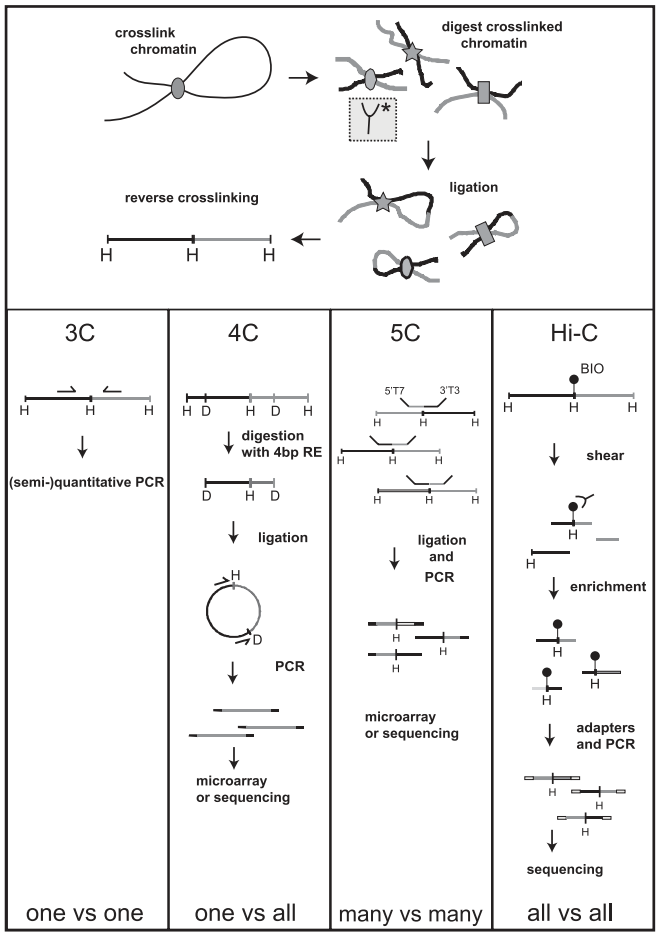

Figure 1 (Adapted from de Wit and de Laat, Genes & Dev., 2012) - Flow chart showing the steps involved in the four major chromosome conformation experiment. First, chromatin is cross-linked using a chemical cross-linker such as formaldehyde. Second, the cross-linked chromatin are then digested with a restriction enzyme. Third, the digested, cross-linked chromatin is ligated under conditions promoting intra-molecular ligation, creating cross-linked ligation circles. Finally, the cross-linking is reversed. Here the four techniques diverge. In 3C, specifically designed probes are used to quantify the proximity probability of two restriction fragments. Therefore, it is referred to as a one vs one technique. 4C probes the proximity probability of one restriction fragment against all other fragments. In 4C, the ligation circles are further digested by another frequent cutting restriction enzyme and then re-ligated. Using outward facing primers designed on the fragment of interest, the ligation circles containing the fragment of interest are linearised using inverse PCR. The linearised products are then sequenced to get proximity probability values between the fragment of interest and all other fragments in the genome. Therefore, 4C is referred to as one vs all. 5C, probes the interaction between many different restriction fragments. To the 3C circles, specific primers are annealed. These primers hybridise to specific restriction fragments. Primers annealed in a head-to-head fashion are ligated by the addition of Taq ligase. This generates the 5C library which is then amplified and sequenced to yield the proximity probability values between all the restriction fragments of interest. Therefore 5C is known as the many vs many procedure. In Hi-C the ligation circles contain a biotinylated base at the ligation junction. These ligation junctions are purified using streptavidin beads after shearing with sonication. These purified regions are then amplified and sequenced to generate proximity probability values between all regions of the genome. Therefore, Hi-C is known as the all-vs-all procedure.

The description of 3C or chromatin conformation capture by Dekker et al., 2002 (Dekker et al. 2002) signalled the beginning of an accelerated growth phase

in the field of higher-order chromatin structure with an active involvement of sequencing technologies. Chromosome conformation capture or 3C (**Figure 1 top, 3C**) as originally described, involves;

- Isolation of intact nuclei.
- Cross-linking of proteins and DNA inside the nucleus by using formaldehyde.
- The cross-linked DNA is then digested with sequences specific/frequent cutting restriction enzymes.
- The cross-linked DNA with restriction enzyme digested ends are then ligated in highly dilute conditions. This promotes intra-molecular ligation. Here, one molecule refers to cross-linked DNA with restriction enzyme digested ends.
- The cross-linking is then reversed, and the ligation products are quantified using qPCR and probes designed for specific ligation products.
- To normalise these values, control ligation products are generated in equal abundance. These regions are quantified and used as a normalisation factor for the cross-linked DNA.

Using 3C, Dekker et al., 2002 (Dekker et al. 2002) were able to recapitulate known general features of the yeast chromatin organization. In brief, they were able to show that the telomeres of chromosomes contacted each other more than expected considering the genomic distance separating them. This was expected, because in yeast the telomeres are known to cluster in 3-D space (Dekker et al. 2002). In premeiotic cells, centromeres in yeast form a cluster near the spindle body. This cluster breaks down during meiosis and is later reconstituted after the first division. They were also able to recapitulate these events. Using the chromosome IV centromere as an anchor (CEN4), they detected strong interactions with the

centromere of chromosome III (CEN3) premeiosis. When probing the same regions interactions after the onset of meiosis, they observed a marked decrease in interaction frequency.

The invention of 3C makes direct probing of regulatory networks possible. More importantly, the actual mechanism of regulation could be quantified. The original 3C protocol utilised HindIII which cuts at sequence specific sites that are 6-bp long, but the protocol can be adapted for a range of different restriction enzymes, such as BglII, SacI, BamHI, EcoRI, AciI, DpnII (de Wit and de Laat 2012). The adaption of 3C with the usage of BglII lead to the first direct evidence showcasing long-range looping interactions during transcription between the murine β-globin LCR and the active globin gene (Tolhuis et al. 2002). Prior to this, evidence supporting looping came from prokaryotic operon systems. The evidence suggested that regulatory sequences separated by large distances and required for the repression of the *gal* and *araBAD* operon were bound by their corresponding repressors (Ptashne 1986). The predominant idea was that proteins at the regulatory sequences interacted with other proteins near the transcription start site and the interjecting DNA looped out from that region. Afterwards, it was shown that the long-range looping interactions in the β-globin locus dynamically change with changes in transcription during development (Palstra et al. 2003) and that these changes are driven by transcription factors (de Wit and de Laat 2012).

3C is limited by the distances and targets that can be probed. Because, regions which are proximal in linear space are also proximal in 3D space an inherent bias is present wherein ligation products from DNA fragments within a few kilobases of the probe-site dominate the sample. Also, 3C only allows the probing of very specific one-on-one interactions. The invention of 4C or Chromosome conformation capture-on-chip (**Figure 1 top, 4C**) was aimed at probing one-vs-all. 4C attempts to

quantify the interaction frequency that any given loci has to interact with the probe site (“viewpoint”). There are two variants of 4C, one uses microarrays containing a preconfigured set of sequences, the other named 4C-seq uses next generation sequencing technologies (Splinter et al. 2011) to capture interactions between the viewpoint and all other restriction sites. There are two main methods of creating 4C libraries.

The first relies on the usage of a single frequent 4-bp cutting restriction enzymes to create cut sites after cross-linking. After de-cross-linking, the ligation circles containing both junctions between the viewpoint and captured fragments are amplified with inverse PCR by using outward facing primers on the viewpoint fragment. The second uses, two restriction enzymes. A 6-bp cutting enzyme is used after cross-linking, this is followed by ligation and de-cross-linking which generates very large ligation circles. Next, a frequent 4-bp cutting restriction enzyme is used to further trim these circles, followed by another step of ligation. Finally, the ligation circles containing two junctions involving the viewpoint fragment are amplified with inverse PCR by using outward facing primers on the viewpoint fragment (Simonis et al. 2007). Notably, 4C was used to highlight the separation between active and inactive regions of the genome using the β-globin gene which is a tissue specific gene against Rad23 a housekeeping gene as a control. It was shown that, Rad23 made contacts with many active regions on its own chromosome and on other chromosomes. But the erythroid specific β-globin gene made contacts with other active regions in erythroid cells. Whereas, in fetal brains the β-globin since it is inactive only contacted other inactive regions (Simonis et al. 2006). The stability of chromosome conformation has also been probed with the help of 4C. Previous FISH studies had suggested that an ectopic human β-globin LCR placed within a cluster of housekeeping genes in mice would move the cluster outside its chromosome

territory (Noordermeer et al. 2008). Later investigating the same scenario with 4C revealed that in reality no new contacts were established (Noordermeer et al. 2011a). 4C based microarrays have also been used to showcase the different chromosome conformations of active vs inactive chromosome X in the context of mammalian dosage compensation (Splinter et al. 2011).

A less sensitive but more specific version of 4C is 5C or chromosome conformation capture carbon copy (**Figure 1 top, 5C**) (Dostie et al. 2006). Rather than probing for a single viewpoint versus all other genomic sites, 5C probes all possible pairwise interactions between a set of predefined viewpoints. First, a 3C library is generated by cross linking, digestion with restriction enzyme, ligation and de-cross-linking. To this library a set of predefined 5C primers originating from the fragments of interest are annealed. Both forward and reverse primers are used. Furthermore, these primers also contain universal PCR primers (T7 for the 5'-ends of forward primers, T3c for 3' ends of reverse primers) at their tails. Next, application of Taq ligase ensures that the annealed 5C primers at the ligation junction are ligated. This creates the 5C library, which captures a part of the 3C library. The final ligation products of interest in this library are the head-to-head 5C primer ligation products between a forward and reverse primer with the universal PCR tails facing outwards. These ligation products are then amplified using the universal PCR primers. Since both forward and reverse primers are present in equimolar quantities, the amplified signal of each head-to-head ligation product reflects the relative enrichment of any given interaction between two genomic loci. 5C can be thought of as being a high-throughput but more specific version of 3C with sensitivity that is lesser than 4C.

5C has been used to address many diverse problems. A detailed structural analysis of the tissue specific α-globin gene in K562 (expressing α-globin) versus

GM12878 (α-globin not expressed) has been done (Baù et al. 2011). Using *in-silco* modelling of a 500 Kb gene dense region of chromosome 16 which harbours the α-globin gene Baù et al. showed, that, in GM12878 cells where α-globin is not expressed this entire region forms a single domain or what the authors call globules, but in K562 cells where α-globin is heavily expressed two such domains are formed. In both globules active genes tend to cluster near the centre, while the inactive genes are positioned towards the periphery of the globule. Most notably, the ENCODE consortium showed that proximity probability values in 5C maps strongly correlated with known regulatory regions the consortium had identified with the help of genome-wide DNase I hypersensitivity screens (Thurman et al. 2012). Furthermore, 5C has also been used to investigate the changes in 3D organisation during cellular differentiation (Phillips-Cremins et al. 2013). Using neural progenitor cells derived from mouse ES cells a diverse set of interactions (90,000 *cis* and 500,000 *trans*) were probed near developmentally regulated genes (*Oct4, Nano, Sox2, Klf4, Nestin, Olig1-Olig2*) at seven different genomic loci. This lead to the characterisation of locus specific higher-order chromosome conformations, cell type specific (ES, NPC specific) and constitutive interactions between different genomic loci (Phillips-Cremins et al. 2013). The activation of proto-oncogenes due to disruption of chromosomal domains (Hnisz et al. 2016) and the structure of the mitotic chromosome (Naumova et al. 2013) has also been investigated using 5C.

Hi-C or high-throughput chromosome conformation capture (**Figure 1 top, Hi-C**) was a technological leap for chromosome conformation capture (Lieberman-Aiden et al. 2009; Belton et al. 2012). Although the technique has evolved considerably over the years, the basic principles underlying Hi-C is still the same as 3C. Coupled with carefully designed statistics, Hi-C allows for the quantification of proximity ligation events between any two genomic loci separated by any given

distance through a single experiment without requiring the usage of pre-designed primers or viewpoints, such as 5C or 4C. Therefore, Hi-C is popularly called the “all-vs-all” C method. Hi-C aims to capture a snapshot representing a subset of the total interaction space. Because Hi-C is extremely scalable, the library complexity is a key factor that affects the quality of a Hi-C experiment. In any given chromosome conformation experiment, restriction fragments which are proximal in linear space are also more probable to be proximal in 3D space, therefore to capture more distant interactions highly heterogeneous (undergoing asynchronous cell division) cell populations are required. In a Hi-C/3C experiments, one cell can contribute one interaction for a given restriction fragment. Therefore, a large population of cells allows for larger library complexity (Belton et al. 2012). After cross-linking and digestion with restriction enzymes, the overhangs are filled in with biotinylated residues for purification of ligation circles. Biotinylated residues do not have very high ligation efficiency therefore un-ligated ends are digested using endonucleases. After the removal of un-ligated biotinylated residues these ligation circles are sonicated. The sonicated fragments containing the ligation junctions with the biotinylated bases are pulled down using streptavidin beads. Finally, these regions are subjected to paired end sequencing (Belton et al. 2012). In the resulting mate-pairs, one mate originates from one restriction fragment whereas the other mate originates from another restriction fragment. Therefore, each such mate pair corresponds to a proximity event between any two restriction fragments. The quantification of these events are not count values, rather these are probabilistic values reflecting the relative probability of any two genomic loci being proximal in 3D space compared to such events occurring in the genome.

Hi-C allowed the first genome-wide view of the chromatin folding landscape. This showed that Hi-C matrices, containing contact probability values could be

partitioned into two separate compartments (Lieberman-Aiden et al. 2009), A and B. Both compartments showcase similar features in the contact space. Genomic loci within these compartments tend to interact more with other genomic loci from the same compartment (A to A or B to B). This happens even when the linear separation between genomic loci from separate compartments (A to B) may be less than their partners in the same compartment (A to A). In general, the contact probabilities between genomic loci in separate compartments tend to be depleted. Compartment A corresponds to active regions based on correlation with, gene-rich regions, higher than average mRNA expression, accessible chromatin and presence of activating or repressing chromatin marks. Compartment B on the other hand corresponds to inactive regions (Lieberman-Aiden et al. 2009). These compartments align with band domain definitions from previous studies (Manuelidis 1990; Bernardi 1995).

These compartments can be further sub-divided into smaller domains, popularly termed as Topologically associated domains (TADs) (Sexton et al. 2012; Dixon et al. 2012). TADs are regions in Hi-C matrices wherein very far apart genomic loci tend to contact each other more than their immediate neighbours.

2.2.1 The different variants of Hi-C

The original protocol of Hi-C as stated above is known as dilution Hi-C (Rao et al. 2014). This protocol has been improved over the years by many different contributors and each is known by the variation it perpetrates. It is important to note, that there are technical and functional modifications in Hi-C. In case of technical modifications, these modifications aim to solely improve the throughput of Hi-C as a technique. Whereas functional modifications aim to improve the the biological context of the read-outs coming from Hi-C. One of the first variants of Hi-C was tethered chromatin conformation (TCC) (Kalhor et al. 2011). Dilution Hi-C relies on cross-linking, digesting and ligating DNA under diluted conditions. Therefore, it

relies on intra-molecular ligation events (between digested DNA ends which are cross-linked to proteins) occurring due to the lower concentration of substrate (cross-linked, digested DNA). Yet, a very high proportion of ligation events occur between random DNA fragments. After cross-linking and restriction enzyme digestion, cysteine residues in cross-linked proteins are tagged with biotin and tethered to streptavidin coated beads (tethering). After tethering the digested 5' overhangs are filled in with biotin tagged bases and ligated. Afterwards, the normal steps in Hi-C are followed. TCC is able to increase the signal-to-noise ratio considerably (Kalhor et al. 2011).

Another technique which attempted to address the issue of signal-to-noise ratio was genome-wide 3C (Duan et al. 2012). This technique is a more high-throughput version of 4C. Herein, using the normal steps of 3C, i.e. cross-linking, digestion with a 6-bp restriction enzyme, inducing intra-molecular ligation, and reversing the cross-links, a normal 3C library is obtained. This 3C library is then further digested with a second 4-bp restriction enzyme, and re-ligated to create even smaller ligation. Now each circle contains one ligation junction each for the 6-bp enzyme and 4-bp enzyme. The ligation junction created by the first enzyme is once again digested and adaptors for EcoP15I are ligated to the cut ends and a biotinylated adaptor is ligated to both these adaptors closing the circle once again. EcoP15I is a type III restriction enzyme, and it makes cut sites 25-30 bases downstream from the recognition sequences. EcoP15I is now used to make these cuts generating the final product containing 25-27 bases from the two restriction fragments on both ends, with the EcoP15I adaptor sequences and the biotin labelled adaptor in the middle. Using streptavidin labelled beads, the products containing the incorporated adaptors are enriched and finally sequenced (Duan et al. 2012). This method was first used to elucidate the principles of 3D genome organisation in yeast

confirming with genome-wide data the Rabl configuration of the interphase chromosomes (Duan et al. 2010).

To address the issue of ligation efficiency in biotinylated bases, Simplified Hi-C was developed. Simplified Hi-C is similar to dilution Hi-C, but eliminates the usage of biotinylated bases and the enrichment step where ligation circles containing biotin are enriched using Streptavidin beads. The underlying principle being, even if these circles are not enriched we should be able to capture the library complexity using higher depth of sequencing. Simplified Hi-C was originally used to elucidate the principles underlying chromatin folding in flies (Sexton et al. 2012).

The restriction enzymes used in Hi-C are generally 4-bp or 6-bp cutters, 4-bp cutters generally produce restriction fragments which have an average size of 256bp. Whereas, 6bp cutters produce restriction fragments which have an average size of 4kb (Simonis et al. 2007). Therefore, Hi-C is unable to go beyond the single fragment resolution using periodically cutting restriction enzymes. This poses a problem for organisms such as *Saccharomyces cerevisiae* which have smaller genes and showcase functionally relevant structures such as gene-loops which range in length between 2 and 10 nucleosomes (Hsieh et al. 2015). To overcome the problem of resolution in smaller genomes and to increase the highest resolution possible, Micro-C was developed. Micro-C makes use of micrococcal nuclease to achieve single nucleosome resolution in Hi-C maps. The technique was originally used to investigate the global 3D organisation of the yeast genome (Hsieh et al. 2015). Due to the increased signal-to-noise ratio, the authors were able to identify chromosomally interacting domains (CIDs) in yeast (regions in Hi-C maps similar to TADs, but inter-chromosomal).

The implications of using formaldehyde based cross-linking was also explored in the development of Micro-C XL (Hsieh et al. 2016). Formaldehyde is a

short cross-linker (less than 2 Å distance between groups), comparatively disuccinimidyl glutarate (DSG) and ethylene glycol bis(succinimidyl succinate, EGS) are long cross-linkers (DSG, 7.7 Å and EGS, 16.1 Å respectively) (Hsieh et al. 2016). Micro-C XL uses a combination of these two cross-linkers to cross-link proteins that are farther apart in 3D space. The usage of these cross-linkers together or in concert with formaldehyde achieves a higher signal-to-noise ratio than using only formaldehyde (Hsieh et al. 2016).

Although, Hi-C allows the probing of all interactions in the genome, it does not allow probing of specific interactions within this set. Capture Hi-C (CHiC) is a method that combines the specificity of 5C with the high sensitivity of Hi-C and allows a user to probe all possible interactions for a given genomic loci of interest (Dryden et al. 2014). This procedure follows the same protocol as Hi-C but incorporates an additional sequence capture step using pre-defined biotinylated long bait RNA. After creating a normal Hi-C library, the library is hybridized with the bait RNA, the biotinylated bait RNAs are then pulled down using streptavidin beads. This step also pulls down any DNA products that were hybridised to it. This particular method has been used to investigate the long-range interactions involving three cancer risk loci implicated in breast cancer using 519 bait regions. The study found long-range interactions occurring between these cancer risk loci and regions surrounding genes implicated in breast cancer (POU5F1, MYC, SOX2, KLF4) (Dryden et al. 2014). Later on, the method was also used to probe long-range interactions involving nearly 22,000 promoters in GM12878 cell lines and CD34+ hematopoietic progenitor cells (Mifsud et al. 2015).

In situ Hi-C (Rao et al. 2014) is the last and most well-known evolution of dilution Hi-C (Lieberman-Aiden et al. 2009). Many groups have tried to address the issue of random ligation events associated with dilution Hi-C using different

approaches (Kalhor et al. 2011; Hsieh et al. 2016; 2015). *In situ* Hi-C attempts to address the same issue by cross-linking while keeping the nucleus intact (Rao et al. 2014).

Previous Hi-C methodologies used SDS to lyse nuclei, deactivate the restriction-enzyme and solubilise the cross-linked protein-DNA network which was then ligated under dilute conditions. But, it was shown using mouse fetal liver cells that less than 15% of DNA is solubilised when using HindIII and about 40% of DNA is solubilised when using MboI for digestion. Furthermore, using the much tested β-globin locus it was shown that the 3C signals are actually generated from the non-solubilised DNA region (Gavrilov et al. 2013). Therefore, the authors concluded that ligation mostly takes place between regions which are already in close proximity within the cross-linked nucleus. Later it was shown using single cells that the removal of this step better preserved the nucleus (Nagano et al. 2015; 2013). This step was incorporated in *in situ* Hi-C resulting in much better signal-to-noise ratio. *In situ* Hi-C has also been modified to achieve even higher resolution using 2-bp cutting restriction enzyme (CviJI) (Darrow et al. 2016). This procedure has been used to interrogate looping interactions involving more than two genomic loci in the mammalian inactivated X chromosome (Darrow et al. 2016).

This list is by no means comprehensive as we make no mention of single-cell methodologies based on Hi-C (Nagano et al. 2013; 2015) or other C technologies lying at the intersection of immuno-precipitation and Hi-C, such as ChIA-PET (de Wit and de Laat 2012) or those that are complementary to Hi-C (Beagrie et al. 2017). Although related, these techniques are out of bounds for the scope of this study and shall not be described.

2.3.0 Analysis of Hi-C data

Hi-C experiments yield as output paired-end reads, which are aligned to the

genome, spurious read pairs are filtered out, and the remaining read pairs are normalised to generate interaction matrices. Interaction matrices are 2 dimensional matrices, with a set of genomic loci on both the x and y axes. The value at any give cell in the matrix corresponds to an interaction frequency between any two genomic loci. For un-normalised matrices, these are the total number of read pairs remaining between the two genomic loci after filtering. For normalised matrices, these are floating point values corresponding to the same read pairs after controlling for experimental and technical biases. For interaction matrices corresponding to the same chromosome, the matrices are symmetric and the contacts themselves are referred to as *cis* contacts. When the interaction matrices are between different chromosomes, the matrices are not symmetric and the contacts themselves are referred to as *trans* contacts. We will briefly cover the topics of Alignment, normalisation and feature detection in Hi-C data as these are the sections most liable to affect downstream analysis of Hi-C data.

2.3.1 Alignment

Sequencing of Hi-C libraries generates paired-end reads, wherein one read maps to one restriction fragment and its mate pair maps to another restriction fragment. In Hi-C data analysis, each mate pair is therefore treated as a single-end read and aligned separately. The type of alignment used contributes towards the overall quality of the analysis being done. Irrespective of the alignment algorithms used, there are two possible ways to align reads in Hi-C analysis.

- A full-read approach, wherein the entire read is aligned to the genome.
- A chimeric approach, wherein each read is aligned in chunks until a unique match is found or until the read cannot be matched any further. This ensures the mapping of reads which span a ligation junction. In a full-read approach, these reads would remain unmapped as the ligation junction is

not present in the genome.

As read-length increases, chimeric read alignment provides higher gains. In our study, wherein we made a detailed analysis of the currently available pipelines and methods in Hi-C data analysis, we noted that as read length increased so did the difference in mapping percentage, chimeric aligners aligned 30.9% more reads when using short reads (36bp) and 55.4% more reads when using long reads (101bp) (Forcato et al. 2017). A positive difference in alignment rates was also observed across aligners (Forcato et al. 2017) compared to full-read mapping with bowtie2 (Langmead and Salzberg 2012). Chimeric STAR (Dobin et al. 2013) in HIPPIE (Hwang et al. 2015) aligned 18.4%, chimeric BWA (Li and Durbin 2009) in HiCCUPS (Rao et al. 2014; Durand et al. 2016) aligned 27.4%, chimeric Bowtie2 (Langmead and Salzberg 2012) in diffHiC (Lun and Smyth 2015) aligned 40.1%.

Furthermore, when aligning long-reads (>100bp) originating from Hi-C experiments using frequent cutting restriction enzyments (CviJI, DpnII, MboI), many reads are multi-mapping reads which map to more than two restriction fragments. Chimeric mapping (Durand et al. 2016; Darrow et al. 2016) coupled with targeted analysis allows the probing of highly complex looping interaction, such as those occurring in different hubs (Darrow et al. 2016).

2.3.2 Normalization

After alignment, each read pair is assigned to their corresponding restriction fragments and various filters (see Materials and Methods) are applied. After filtering, each read pair corresponds to a count representing an interaction event between two restriction fragments. These restriction fragments are summarised into bins of fixed genomic length to increase the statistical power of the analysis. Since Hi-C produces a population-average map showcasing a subset of all possible interactions, the farther apart two restriction fragments are in linear space, the more

rare is the interaction between them. Resolution gains from Hi-C are predicted to increase as the square root of sequencing depth (Lieberman-Aiden et al. 2009; Mifsud et al. 2015). Thus, restriction fragments are aggregated into equally sized bins and their corresponding counts contributes towards the total count observed between their bins.

Hi-C experiments also have associated biases. These biases are both experimental and technical. Broadly, procedures for modelling and controlling these biases fall under two distinct categories; the explicit procedures and the implicit procedures (Ay et al. 2014; Forcato et al. 2017). Explicit procedures attempt to compute a normalization factor by modelling for known biases such as GC content and mappability, which are two of the major biases affecting Hi-C data (Yaffe and Tanay 2011). Yet, explicit procedures are not able to control for biases such as restriction enzyme efficiencies and cross-linking efficiencies (Ay et al. 2014). Implicit procedures on the other hand control for unknown biases and are based on the assumption that every single loci is equally probable to interact with every other genomic loci. The most famous of these methods is ICE (Iterative Correction and Eigenvector decomposition) (Imakaev et al. 2012).

The original study which introduced Hi-C purported a simplistic coverage based normalization factor for Hi-C data (Lieberman-Aiden et al. 2009), called “Vanilla coverage”. Vanilla coverage computes a multiplicative normalization factor from the reciprocal row sums and col sums. Then each cell corresponding to an interaction value between two genomic loci (row and column) is normalised by multiplying with the reciprocal of the row sums and the reciprocal of the col sums.

Two explicit normalization factors have originally been proposed. This procedure computed an expected multiplicative factor based on the mappability, restriction fragment length and GC content of a genome digested by any given

restriction enzyme (Yaffe and Tanay 2011). This procedure was further extended later on for high-depth sequencing with modifications that included modelling factors related to fragment length and distance together whilst factors related to GC bias was modelled separately. Furthermore, the original explicit bias modelling procedure binarised the interaction matrix between restriction fragments such that every single interaction was quantitatively equivalent to any other interaction. In the newer procedure, this particular step was removed (Jin et al. 2013). The original explicit biases modelling procedure was also extended to HiCNorm (Hu et al. 2012). Herein, the fragment length and GC content features are estimated, whereas the mappability feature is treated as a Poisson offset.

Another study proposed GC content and fragment length as biases that affect Hi-C data, but they additionally proposed circularisation of ligation products as a bias that affects Hi-C data. To account for these biases an implicit procedure, Sequential Component Normalization (SCN) was proposed (Cournac et al. 2012), wherein using euclidean normalization, separately the rows and columns of a matrix are normalised to 1 until convergence is achieved.

Implicit normalisation factors as stated earlier attempt to control for unknown biases. The first such method was the Iterative Correction and eigenvector decomposition procedure (ICE) (Imakaev et al. 2012). ICE, like other matrix balancing algorithms that have been proposed after it, attempts to compute the normalisation factor for the rows and columns of an interaction matrix separately, such that the variance between each cell in the row or the column is minimised beyond a certain threshold. ICE starts off from an initial bias vector corresponding to the mean observed interaction frequency for every given row or column and uses it to seed the first iteration of the procedure. It then estimates the bias by using the single-sided reads (where only one mate pair maps to the genome) and attempts to

minimise the variance between all cells in either the row or the column of an interaction matrix. Another popular normalisation using matrix-balancing algorithms is the Knight Ruitz procedure. Similar to ICE, it is noted to be much faster (Rao et al. 2014).

As resolution of Hi-C datasets have increased over the years, a new class of analysis procedures have been proposed. These are the bin-less or bin-free approaches to Hi-C data analysis (Spill et al. 2017; Cohen et al. 2017). SHAMAN, the first of its kind uses a Markov Chain Monte Carlo randomisation approach to randomise contact distribution such that the genomic distance between contacts is preserved and the marginal contact distribution is also preserved (Cohen et al. 2017). Finally, to check for the enrichment of Hi-C contacts around any given region, rather than taking binned regions, a bin free comparison is achieved by finding all contacts that are within a certain distance of the given region of interest in both the observed and the randomised matrices. The normalised score of these contacts is the Kolmogorov-Smirnov D statistic obtained by comparing the observed and randomised matrices. This procedure has been applied towards the analysis of the highest resolution Hi-C analysis during mouse neural development (Bonev et al. 2017). Binless (Spill et al. 2017), another bin-free normalization procedure, proposes a normalization procedure at the fragment level, this is sharp contrast to current methodologies which propose normalizations on binned matrices. Binless estimates biases from the discarded fraction of read pairs using Generalized Additive model fitting (Spill et al. 2017), which uses a negative binomial fit to estimate the normalization factors (Spill et al. 2017). This procedure is similar to iterative correction (Imakaev et al. 2012) but does not make the assumption of equal probability of interaction between all loci. Unlike ICE, this allows rows and columns in a matrix to deviate from the mean value (Spill et al. 2017). Read pairs are

normalised only once, post-normalisation these read pairs are the summarised at the bin level (Spill et al. 2017).

Previously, it has been shown that the choice of normalisation does not affect the inferences made in binned Hi-C data (Rao et al. 2014). Although, this cannot be considered true when considering high-resolution Hi-C matrices (near fragment level resolution) where the assumptions made on low-resolution Hi-C matrices start to break down. Also, this same statement does not apply to the bin-free class of normalizations, since these methods represent a paradigm shift in how Hi-C data is analysed.

2.3.3 Feature detection in Hi-C data

Hi-C data imparts two types of structural information, the first relates to the structural compartmentalisation of the genome. Whereas, the second relates to the detection of long-range looping interactions such as those involving enhancers and promoters, as evidenced by numerous experiments on the α and β-globin locus (Dhar et al. 1989; Lois et al. 1990; Baù et al. 2011; Tolhuis et al. 2002; de Wit and de Laat 2012).

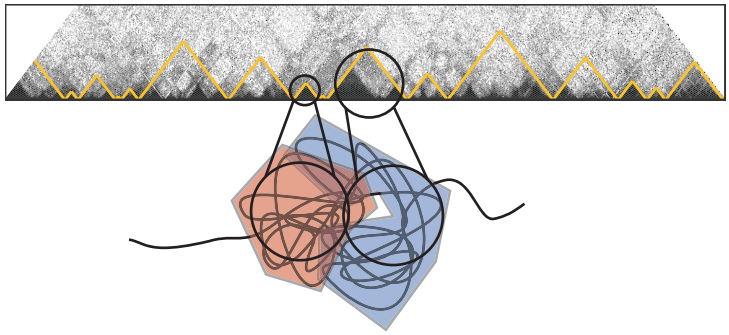

Figure 2 - TADs or Topologically Associated domains are regions of aggregated chromatin. Shown is a cartoon of what two TADs may look like (bottom). Read pairs originating from within these aggregates tend to be over represented in the Hi-C maps and appear as dark triangles (top). The overlapping region between the two chromatin aggregates (bottom) corresponds to the inter-TAD or proximity probability values between two TADs.

Depending on the length scale of the analysis, genome compartmentalisation is referred to as compartments or TADs. Compartments are the first level of genome

compartmentalisation. Compartment A correspond to active and open chromatin, whereas compartment B corresponds to inactive and closed chromatin (Lieberman-Aiden et al. 2009). Both compartment classes show a high-degree of within class clustering, yet do not showcase a great degree of cross-talk. Since compartments correspond to active and inactive regions, these regions change when moving between cell-types and tissue-types (Dekker and Heard 2015). A further level of organisation within compartments are topologically associated domains or TADs. TADs are regions in Hi-C maps, representing highly dense regions of interaction between distant genomic loci (**Figure 2**). TADs have been described in a range of sizes, starting from a few kilobases upwards to several mega bases in mammals (Nora et al. 2012; Dixon et al. 2012). Smaller scale TADs have been described in flies (Sexton et al. 2012; Hou et al. 2012). Co-regulated genes tend to occur in the same TAD (Le Dily et al. 2014). TAD boundaries also play host to a number of architectural proteins such as CTCF in mammals (Dixon et al. 2015; 2012; Phillips-Cremins et al. 2013) or BEAF-32 and CP190 in flies (Sexton et al. 2012; Hou et al. 2012) and are enriched in their binding sequences (Ramírez et al. 2018). TADs are also known to have a hierarchical organization, with each TAD being partitioned into smaller TADs (Phillips-Cremins et al. 2013; Berlivet et al. 2013).

The existence of TADs, their invariant nature and the functional implications of these structures on genome regulation has been a point of major investigation. TADs correlate with early or late replicating regions and harbour entire regions showcasing differential replication timing from those regions that showcase uniform replication timing (Pope et al. 2014). Transcriptional states are also predictors of TAD structuring (Rowley et al. 2017). Furthermore, in *Drosophila* the onset of transcription during development coincides with the appearance of TADs (Hug et al. 2017). TADs are also known to be conserved structures across prokarya (Dekker

and Heard 2015) and eukarya (Rudan et al. 2015). Directionally oriented CTCF binding motifs across the genome are thought to act as barrier that regulate the direction of long-range looping interactions (Rao et al. 2014; Rudan et al. 2015). TADs are also invariant during differentiation (Nora et al. 2012; Dixon et al. 2012), while the TADs themselves don't change they showcase an increase or decrease of contact frequency within TADs (Dixon et al. 2015). Multiple studies have attempted to investigated the effect of TAD boundary disruption. Most notably, studies in the HoxD locus have showcased that TAD boundaries are highly resilient to change and only very large deletions lead to the merging of TADs separated by the deleted boundary (Rodríguez-Carballo et al. 2017). This resilience was also showcased previously when a small deletions (35Kb) in the HoxD locus lead to an increase in the expression HoxD11, but a smaller deletion had no effect (Noordermeer et al. 2011b). The re-composition of TADs after such large-scale re-arrangements if induced are mediated by the re-hashing of existing CTCF sites (Fabre et al. 2017). CRISPR mediated inversions in the CTCF binding site in the protocadherin (*Pcdh*) cluster (Guo et al. 2015) have also showcased how CTCF directionality mediates long-range looping interactions. Simulations have also postulated that supercoiling induced plectoneme formation may also play a role in affecting Cohesin mediated loop-extrusion (Fudenberg et al. 2016) and the formation of TADs (Racko et al. 2017). CRISPR induced genomic rearrangements mimicking deletion, inversion or duplication of CTCF binding sites in the TAD harbouring WNT6/IHH/EPHA4/PAX3 loci have also linked such changes to limb malformation (Lupiáñez et al. 2015). The causative link between gliomas and IDH mutation has been studied from the perspective of higher-order chromatin structure. It has been observed that IDH mutant cells gained methylation in nearby CTCF sites leading to lower insulation and the increase in contacts between PDGFRA, a cancer

driver, and enhancers outside the PDGFRA TAD (Flavahan et al. 2016).

The identification of TADs and long-range looping interactions is primarily done using bioinformatic algorithms. Many such algorithms have been proposed for TADs(Filippova et al. 2014; Serra et al. 2016; Dixon et al. 2012; Haddad et al. 2017; Zhan et al. 2017; Crane et al. 2015) and loops(Rao et al. 2014; Lun and Smyth 2015; Hwang et al. 2015; Ay et al. 2014; Mifsud et al. 2017). In our recent study, where we conducted a comparative assessment between several TAD callers and loop-callers, we found that on a general basis TAD calling algorithms had very high concordance between replicates, but the same is not true for loop calling algorithms. Furthermore, loop calling algorithms showcased a strong dependency between the number of loops called and the total coverage of the dataset (Forcato et al. 2017).

2.4.0 Dosage compensation in *Drosophila melanogaster*

The process of dosage compensation is a highly plastic phenomenon that affects a change in transcriptional regulation, balancing the transcriptional output originating from sex chromosomes between males and females in species where a copy number difference exists between males and females (Ferrari et al. 2014; Samata and Akhtar 2018). Many different models of dosage compensation are known. The three most well-known models are that of placental mammals, *C.elegans*, and *Drosophila*. In placental mammals, dosage compensation silences an entire X chromosome in females achieving equivalency between males and females. In *C.elegans* the expression of the X chromosomes in the hermaphrodite is halved. In *Drosophila*, the male X chromosome is up-regulated by two-fold. Although, all of these models operate in a dissimilar fashion, they follow a very similar pattern. The dosage compensation complex is first recruited at nucleation sites, wherefrom it spreads across the chromosome and affects a change in transcriptional response (Ferrari et al. 2014). While there exists a plethora of

questions related to the molecular aspects of dosage compensation, my interest is skews towards the chromatin structural aspects of dosage compensation in *Drosophila*, where only active genes are up-regulated by a non-constant factor and in-active genes remain silent. This up-regulation is such that a genome-wide average of 2x up-regulation is achieved (Ferrari et al. 2013). The inactivation of the mammalian X chromosome starts at the X-inactivation centre, from the X-inactivation centre it spreads across the X chromosome and affects silencing. Although a few genes are still missed, most of the inactive X chromosome adopts a distinct structure, devoid of TADs and partitioned in the centre (Giorgetti et al. 2016). In *C.elegans*, the hermaphrodite X chromosomes are down-regulated such that gene expression is halved. In this case, the X chromosome shows a change in insulation that is very different from the autosomes (Crane et al. 2015).

In fly, non-coding RNAs, *roX1* and *roX2* in addition to other proteins (MSL1, MSL2, MSL3, MOF and MLE) comprise the dosage compensation complex. MSL1 is the scaffold protein which holds the entire complex in place. MSL1 interacts with MSL2 via a coiled-coil domain in its N-terminus and with MSL3 and MOF via a PEHE motif in the C-terminus region (Samata and Akhtar 2018). As previously stated, the MSL2 protein is repressed by Sxl in females. MSL2 is expressed in males and MSL2 together with roX2 ncRNA allows the sequence specific targeting of the dosage compensation complex (Samata and Akhtar 2018). MSL3 facilitates the spreading of the dosage compensation complex across gene bodies and strengthens the acetylation activity of MOF. MOF carries the histone acetyl (H4K16Ac) transferase activity. This particular activity has been linked to increased chromatin decompaction and enhanced transcriptional output. Although, the exact mechanism by which this is achieved is highly debated. MLE is a helicase linked to the unwinding of chromatin and is responsible for effective loading of roX2 non-coding

RNAs (Samata and Akhtar 2018).

The dosage compensation complex is initially recruited to sequence-specific sites on the X chromosome known as high-affinity sites (HAS) (Straub et al. 2008) or chromatin entry sites (CES) (Alekseyenko et al. 2008). Different lists have been identified using different techniques. Hereon, we will refer to these sites interchangeably as dosage compensation binding sites or MSL binding sites. Dosage compensation binding sites are known to be present near transcriptionally active genes. There is also a correlation between the distance to the nearest binding site and the transcriptional output of the gene (Samata and Akhtar 2018). The dosage compensation binding sites contain a 21bp GAGA rich motif, called the MSL recognition element (MRE) (Alekseyenko et al. 2008). These motifs are present in autosomes but are not recruited there. Therefore, it was postulated that the higher-order chromatin structure had a role to play in the recruitment of the dosage compensation machinery to its effector sites. Recently, it was shown that PionX sites (Villa et al. 2016), a subset of MSL binding sites provides sequence specificity for early establishment of MSL binding (Schauer et al. 2017). A second protein, chromatin-linked adapter for MSL proteins (CLAMP) has been previously shown to bind the MRE in *drosophila* and is also involved in the recruitment of the dosage compensation complex to the X chromosome (Soruco et al. 2013). Furthermore, CLAMP binding also creates very large regions of open chromatin near its binding sites (Urban et al. 2017). The CLAMP protein binds the genome non-specifically but has the highest binding signal at regions that are bound by MSL (Soruco et al. 2013). Based on this, a subset of CLAMP binding sites were categorised as MSL dependent, partially MSL dependent and MSL independent binding sites (Soruco et al. 2013). Finally, it has also been shown that dosage compensation binding sites tend to colocalize in three-dimensional space aiding the spreading of the dosage

compensation complex to progressively more inaccessible regions of the genome (Ramírez et al. 2015).

The structural changes that accompany dosage compensation have been extensively studied in *c.elegans* and mammals. The mammalian dosage compensated X chromosome in females adopts a distinct structure. An increase in insulation in the *C. elegans* hermaphrodite X chromosomes has been reported post-DC. In the *drosophila* dosage compensated X chromosome, a changed structure of the X chromosome was previously postulated (Grimaud and Becker 2009) based on FISH experiments, yet recent studies using Hi-C were unable to detect these changes (Ramírez et al. 2015; Schauer et al. 2017). This is partially due to the inherent problems that accompany dosage compensation in *Drosophila*. Previous studies were done using cell lines. In *Drosophila* cell lines, the male S2 cell line and the female Kc167 cell lines are biased by copy number differences. Specifically, the female Kc cells are on average tetraploid (Lee et al. 2014). The S2 cells also carry several copy number changes (Lee et al. 2014). These copy number differences may hinder downstream analysis of Hi-C data as the assumption of equal visibility of all genomic loci does not hold true for the more popularly used implicit normalization methods (Imakaev et al. 2012). In the wild-type, *drosophila* males carry a single X chromosome. This, compared to the female two X chromosomes ensures that at equal sequencing depth, the male X chromosome has half as many reads as the female X chromosomes. To ensure that over-correction does not occur for the single X chromosome and to not be biased by the inherent copy number bias present in *drosophila* cell lines, we used high-resolution Hi-C data generated using sex-sorted embryos and adopted chromosome specific normalisation procedures (Ramírez et al. 2015).

We have previously demonstrated that peak callers are positively correlated

to the sequencing depth of the experiment (Forcato et al. 2017). To ensure an equivalent comparison between the male and female samples I devised a non-parametric procedure for comparing highly interacting regions of the genome. Finally, I have also devised a novel, TAD boundary calling procedure that is both fast and accurate and is sensitive to small scale domains on the chromatin fibre. Using these tools and carefully designed analysis procedures we were able to detect previously unknown differences in the male dosage compensated X chromosome.

3.0 Materials & Methods

The passages herein have been quoted verbatim or adapted from the following sources: Pal et al., (manuscript in revision).

Hi-C data processing

Hi-C data was processed with the hiclib (2016-07-14 version - commit fe3817a; https://bitbucket.org/mirnylab/hiclib) and cooler (v0.3.0; https://github.com/mirnylab/cooler) packages by Leonid Mirny's lab for ICE normalization(Imakaev et al. 2012). hicpipe based explicit normalisation was also applied to specific cases as an alternative. hicpipe was used for the probabilistic bias modeling normalization proposed by Yaffe and Tanay (Yaffe and Tanay 2011). Whereas, ICE was used for implicit matrix balancing normalisation.

We aligned reads to the dm3 genome build considering only chrX, 2 and 3. chr4, Y and the heterochromatic portions (named with suffix "Het") were left out. For ICE, bowtie2 was used for alignment. Whereas, hicpipe used bowtie (Langmead et al. 2009) (v1.1.2) for alignment.

For hicpipe we used default parameters, except SEGMENT_LEN_THRESHOLD, which was set to 800 for the sex-sorted embryos dataset. This parameter was set after examining the distribution of the sum of distances between read pairs and their nearest downstream fragment end.

For the ICE pipeline the iterative_mapping module in hiclib was used for aligning reads to the reference genome. hiclib alignments were run using Bowtie2 (Langmead and Salzberg 2012) version 2.2.9. For the sex-sorted embryo datasets (GSE94115) the following parameters were adopted: min_seq_len=20, len_step=10, seq_start=0 and seq_end=49. In the S2 and clone-8 cell lines data obtained from Ramirez *et al.* (Ramírez et al. 2015) we used: min_seq_len=20,

len_step=10, seq_start=0 and seq_end=50. For the Kc167 cell line data from Li *et al.* (Li et al. 2015) we used: min_seq_len=20, len_step=10, seq_start=0 and seq_end=50. Additional bowtie2 flags were --mm and --very-sensitive.

The following filtering parameters were applied for hiclib: For embryos, S2 and clone-8 samples the maximumMoleculeLength was set to 800, for Kc167 samples maximumMoleculeLength was set to 300 (as in the original publication). Duplicates were filtered using the filterDuplicates function. Later, the technical replicates were merged into their corresponding sample. The final read numbers are available in Table 1.

The Hi-C data has been summarised at several resolutions (bin sizes), including 25Kb, 10Kb and 3.5Kb. At the highest resolution (3.5Kb bins) we verified that in the Hi-C maps at least 80% of the bins had at least 1000 reads as proposed previously (Rao et al. 2014). Finally, the binned matrices were normalised with ICE chromosome by chromosome (chromosome-wise) using mirnylib.numutils.iterativeCorrection and genome-wide using cooler iterative_correction. To allow rows or columns for normalization we required at least 40 as sum of read counts. Furthermore, to remove non-informative read pairs the first two diagonals were removed during normalization (interactions at distances 0 or 1 bin). Finally, the tolerance value was set to 1e-02.

Computing decay of Hi-C signal

The interactions at distances ranging from 2 bins (50Kb for 25Kb matrices) to 100 bins (2.5Mb for 25Kb matrices) were considered. In the normalized Hi-C matrices, NAs, NaNs and infinite values were set to 0. The median Hi-C signal (y-axis) was computed at each distance (x-axis).

When indicated, the Hi-C signal was transformed into contact probabilities (contact

frequencies) by assuming the contact probability is maximum (equal to 1) when considering neighboring genomic loci. To this concern the median normalized Hi-C signal is computed for each diagonal and divided by the median signal at the first informative diagonal (2 bins distance) to obtain contact frequencies. Then the median contact frequencies are $\log_{10}$ transformed (y-axis) to be plot against the log of genomic distance (x-axis) in the log-log plots. This procedure is applied in Figures 5, 7, 8.

Previous literature proposed an alternative probabilistic transformation of Hi-C matrices (Giorgetti et al. 2014), based on the same assumptions of maximum contact probability near the diagonal. We also applied this transformation where the signal inside every cell of the Hi-C matrix is divided by the mean normalized signal at the first informative diagonal (2 bins distance) to obtain a contact probability. Any resulting value greater than 1 was set to 1. This method is only applied to Figure 11.

We then used the lm function in R to fit a linear model to the values in the log-log plot to obtain the slope coefficient. The linear model fitting was done for values at distances ranging from 2 bins (50Kb or 4.69 in the $\log_{10}$ scale) to 15 bins (375KB or 5.57 in the log scale), i.e. in a range of distances where the decay is close to linear in the log-log plot.

The interaction decay differences are assessed by computing the pairwise differences of slope coefficients (deltas) between autosomes or between chrX and autosomomes. The slope coefficient deltas of chrX vs autosomes are then compared to those between autosome pairs using Wilcoxon test as indicated in individual boxplots.

Alternatively, to assess the difference between the interaction frequency plots, the cumulative density functions (CDFs) of the interaction probability for autosomes or chrX are computed. CDFs of interaction probability were estimated from 50Kb to

2.5Mb as cumulative sums of median Hi-C contact frequencies for each distance, then divided by the cumulative sum maximum value to make it equal to probability 1. Kuiper's statistic for pairwise comparisons between autosomes or between chrX and autosomomes is then computed as the sum of absolute values for the maximum positive and negative differences between CDFs as

$$\mathrm{V} = \max\left(cdf_{chr1} - cdf_{chr2}\right) + \max\left(cdf_{chr2} - cdf_{chr1}\right)$$

The difference in the estimated pairwise Kuiper's statistics of chrX vs autosomes are then compared to those between autosome pairs using Wilcoxon test as indicated in individual figures.

Down-sampling of Hi-C matrices

To account for the disparities in coverage due to copy number and sequencing depth differences between male and female samples, we used two approaches for down sampling of read counts, as indicated in the text. In the first, we simulated the effect of single copy number on male autosomes by randomly down sampling 50% of the autosomal reads in the male samples. For this we used the numpy.random.binomial function in python with the probability parameter set to 0.5. The down-sampled observed read counts were then normalized using chromosome-wise ICE. This method was used to check the rate of Hi-C decay when the copy number of autosomes is similar to that of chrX (Figure 7).

In the second approach, we down-sampled all female chromosomes (observed cis read counts) by the ratio of cis interactions count present in the corresponding male chromosome, to make the total sum of observed cis read counts comparable between male and female samples, chromosome by chromosome. The down-

sampled observed read counts were normalized with chromosome-wise ICE. This approach was used to verify the effect on TAD calls, and the effect on clustering of top-scoring interactions between the male and down-sampled female samples (Figure 19, Figure 29).

Polymer folding simulation

We simulated the generic large-scale dynamical folding of the diploid Drosophila genome using Rigid-body Langevin Dynamics (Carrivain et al. 2014) at room temperature T=300 K. The eight chromosomes were modelled as simple self-avoiding polymers composed of 10-Kb segments (rigid cylinders of length 170 nm and diameter 25 nm corresponding to a 10-nm fiber). At the beginning of each simulation, chromosomes started in a mitotic Rabl-like configuration, followed by a smooth confinement into a sphere of diameter 4 μm mimicking the nucleus. Then the dynamics of the genome was tracked during two hours of real time. Average contact probabilities were calculated over thousands of independent simulations. As in Hi-C, we merged *trans* contacts between homologous with *cis* intra-chromosomal contacts.

We considered four situations: two with a female diploid genome (two copies of chromosomes 2, 3, 4 and X) and two with a male genome (two copies of chr2, 3, 4 and one copy of chrX and Y). For both sexes, we examined one case without pairing between homologous chromosomes and one case where pairing was imposed by adding springs between homologous segments every 100Kb.

Non-parametric selection and clustering of top-scoring interactions

For the non-parametric selection of top-scoring interactions we used normalized Hi-C data binned at 25Kb bins. NAs, NaNs and Infinite values were set to 0 and we discarded the first two diagonals (interactions occurring at distances 0 or 1 bin). We

then selected the highest 5% (default threshold, applied unless otherwise specified) of normalized Hi-C contact values in any given diagonal as the top-scoring interactions (Figure 16). When indicated, different thresholds were adopted as percentage of highest scoring interactions, as well as thresholds on the maximum distance of interacting loci pairs.

To define clustered top-scoring interactions we consider the euclidean distance between any pair of top-scoring interactions (*i, j*) with coordinates (i_x, i_y) and (j_x, j_y), respectively, in the space of Hi-C matrix bins coordinates. With bin size 25Kb, the distance *D* for each pair is defined as:

$$D = \left(\sqrt[2]{(i_x - j_x)^2 + \left(i_y - j_y\right)^2} \right) \times 25000$$

If distance $D \leq 25Kb$ interaction j and i are grouped under the same cluster name. During merging, in an iterative process the list of clusters is scanned and clusters sharing elements are merged into larger clusters. Finally, we obtain a list of clusters containing unique interactions. We report the difference in the proportion of clustered top-scoring points. With default settings ($D \leq 25Kb$) the procedure is equivalent to cluster neighboring top-scoring interactions only.

Estimating propensity of each chromosome to participate in *trans* interaction

For each chromosome pair (a, b), where $a \neq b$ are chromosomes $\{2L, 2R, 3L, 3R, X\}$ the expected number of trans interactions is estimated with a null model where trans interactions originating from any chromosome are uniformly distributed over the other chromosomes (targets). This is estimated by adjusting the expected counts by the target chromosomes length and copy number. For example, the expected

trans interactions $E_{2L,2R}$ originating from chr2L and targeting chr2R is estimated as:

$$E_{2L,2R} = \frac{(c_{2R} \times l_{2R})}{\sum(c_b \times l_b)} \times T_{2L}$$

where b contains the set of target chromosomes $\{2R, 3L, 3R, X\}$ (all except the origin chromosome 2L). Whereas c_i and l_i are the expected copy number and length, respectively, of the specified chromosome *i*. Then T_{2L} is the total number of trans contacts originating from the chromosome 2L.

Defining domain boundaries in 3.5Kb bins using LSD

Domain boundaries have been defined on 3.5Kb bins matrices using Local Score Differentiator (LSD)(code available at https://bitbucket.org/koustavpal1988/fly_dc_structuralchanges_2018/). The directionality index (DI values) was computed as in Dixon et al. 2012 (Dixon et al. 2012) on a window of 35Kb (10 bins) using the *ComputeDirectionalityIndex* function. We then computed the forward and backward differences of the DIs using the *Forwards.Difference* and *Backwards.Difference* functions defined as the difference in DIs between a bin and its adjacent downstream or upstream bin, respectively.

$$\Delta DI_{forward} = DI_i - DI_{i+1}$$

$$\Delta DI_{backward} = DI_i - DI_{i-1}$$

We then identify domain starts and domain ends using the outliers of the forward and backward differences within a local window of 30 bins corresponding to 105kb

in a 3.5Kb binned matrix. Outliers are detected as follows:

First, we define fences on the forward and backward differences distribution as

$$Fence_{forward} = Q\left(\Delta DI_{forward}, 0.25\right) - 1.5 \times (Q\left(\Delta DI_{forward}, 0.75\right) - Q\left(\Delta DI_{forward}, 0.25\right))$$

$$Fence_{backward} = Q(\Delta DI_{backward}, 0.75) + 1.5 \times (Q(\Delta DI_{backward}, 0.75) - Q(\Delta DI_{backward}, 0.25))$$

where, $Q(\Delta DI, 0.75) - Q(\Delta DI, 0.25)$ is the interquartile range ΔDI, $Q(\Delta DI, 0.25)$ and $Q(\Delta DI, 0.75)$ are the 25th and 75th quantiles of the ΔDI distributions within the window. 1.5 is the *Tukey's constant* used to select outliers in the local window values distribution.

Domain starts require the DI value to be finite, $\Delta DI_{forward} \leq Fence_{forward}$ and $\Delta DI_{forward} \leq DI$. Domain ends require the DI value to be finite, $\Delta DI_{backward} \geq Fence_{backward}$ and $\Delta DI_{backward} \geq DI$.

An additional filter, requiring $DI \leq 0$ for domain starts, and $DI \geq 0$ for domain ends is also applied for a stricter definition of boundaries (*strict* parameter). This parameter was set to FALSE (*strict*=FALSE) in the analyses for this study, unless otherwise noted. LSD by default also attempts to fill in any gaps that may exist between two called domains by connecting the end and start of two consecutive

domains (*Fill.gaps* parameter), this parameter was set to FALSE (*Fill.gaps*=FALSE) in the analyses for this study, unless otherwise noted.

As LSD identifies domain starts and ends separately, a list of unique domain end positions is considered and extended on both sides by 1/2 bin size to obtain bins spanning adjacent start and end bins as reference border region for downstream analyses. We used the *MakeBoundaries* function to carry out this transformation and obtain 3.5Kb (equal to bin size) wide domain border regions.

Defining domain boundaries using other TAD callers

Armatus (Filippova et al. 2014) (v2.1) TAD caller was obtained from https://github.com/kingsfordgroup/armatus, and run with the parameters $-r$ specifying the resolution (10Kb), $-g$ specifying gamma values ranging from 0.1 to 1 with 0.1 step $\{.1, .2, \ldots, 1\}$ and $-m$.

DomainCaller (Dixon et al. 2012) was obtained from the public repository by the original authors (http://bioinformatics-renlab.ucsd.edu/collaborations/sid/domaincall_software.zip) and was run with directionality index computed at 2Mb distance on 10Kb matrices. As previously reported by multiple groups (Rao et al. 2014; Forcato et al. 2017) the original code was affected by a problem causing the program to exit due to a division by zero in random generated numbers that may occur randomly with larger matrices. To circumvent this problem we used the patch as proposed in(Forcato et al. 2017), where the program reiterates the random number generation.

TADBit (Serra et al. 2017) (v0.1_alpha.360)(https://github.com/3DGenomes/TADbit) was executed using default parameters on uncorrected counts in 10Kb bins matrices.

In all three cases, we computed the proportion of non-matching domain boundaries

using as reference the list of TAD starts produced by the TAD callers.

Defining boundary change annotations

We used exact match of domain boundaries, i.e. intersection of the lists of genomic bins marking the boundary, to classify boundaries as disappearing, appearing or unchanged between the male and female samples.

Insulators binding at domain boundaries

Insulators binding peaks obained from ChIP-chip experiments were first queried on modMine and downloaded from the modENCODE data repository(Contrino et al. 2012). In particular, we used BEAF32, CP190 and CTCF in Kc167 (respective IDs: 3745, 3748, 908), in S2 (respective IDs: 274, 925, 3281) and BEAF32, CP190 and CTCF in embryos (5130, 5131, 5057, respectively).

The binding peaks were overlapped to the 3.5Kb binning table associated with the chromosome-wise ICE normalized Hi-C matrices using the GenomicRanges package(Lawrence et al. 2013). The number of overlaps per bin was counted for each peak file using *countOverlaps* function.

We then created a 10 bin (35Kb) window around the domain boundaries. To do so, we considered the bins mid-point as reference coordinate. Boundaries at less than 35Kb distance from the start and end of the chromosome were removed. Then we aggregated the peaks count per bin for each insulator and boundary class, and the counts were averaged. Finally, for visualization we applied spline smoothing as implemented in the ggplot package (geom smooth, glm method with natural cubic spline and 10 degrees of freedom).

To compute the median insulators enrichment around domain boundaries, we used the same ChIP-chip datasets listed above, for which we retrieved the enrichment

signal (.wig) files from the modENCODE data repository. Signal files (.wig) were rescaled by dividing the signals in each file by their 99th percentile, to facilitate comparisons across datasets accounting for potential differences in ChIP efficiency. Insulator average profiles were calculated using deepTools (Ramírez et al. 2016) (version 2.5.3). Each average profile is displayed in a 10Kb window centered on domain boundaries, with a bin size of 100bp.

Dosage compensated gene annotation

The list of genes responding to dosage compensation were obtained from Zhang et al., 2010 (Zhang et al. 2010) (GEO GSE16344). Following their criteria, we considered genes detected in all replicates, then

selected genes with mean expression $\geq$4 RPKM in wild type control S2 cells and ratio $\leq$ 0.74 between mean expression after MSL2 knockdown vs control.

MSL binding sites definition

MSL binding site definitions were obtained from three previous articles. The refined list of High Affinity sites (HAS) were obtained from Ramirez et al. 2015 Table S2 (Ramírez et al. 2015), and the original HAS list was obtained from Straub et al. 2008 Table S1(Straub et al. 2008). CES sites were obtained from Alekseyenko et al. 2008 Table S1(Alekseyenko et al. 2008).

CLAMP binding sites definition

CLAMP binding sites as defined by Soruco et al. 2013(Soruco et al. 2013) were provided by E. Larschan.

Computing enrichment of MSL binding sites and CLAMP binding sites around domain boundaries

Mid-points of MSL and CLAMP binding sites were used as reference positions. For each factor (*m*) We computed the randomly expected binding sites per genomic bins (E_m) assuming a uniform distribution as null model: i.e. we divided the total number of binding sites (N_m) by the length of chrX (L_X) measured as number of (3.5Kb) bins.

$$E_m = \frac{N_m}{(L_X)}$$

Next, for each domain border (belonging to the disappearing, appearing or same classes), we considered a window with size up to 15 bins (52.5Kb) on both sides. If such windows overlap for any pair of neighboring domain boundaries, they are shortened by assigning equally to both boundaries the intervening region. This is an important point as avoids overestimating the association of any boundary class to genomic features, while allowing at the same time a definition of boundaries at fine scale (i.e. small domains).

Then we counted the number of binding sites windows around boundaries of each class, then divided by the windows total length. This result is the observed average number of binding sites per bins in the regions around boundaries of each class. The final results are reported as $\log_2$ ratio of observed over expected average number of binding sites per bin.

Computing distance of dosage compensated genes to nearest domain boundary

Around domain boundaries we considered a window of up to 15 bins, adjusted for

overlap with neighboring boundaries windows as described above. We then used the *findOverlaps* function from the GenomicRanges(Lawrence et al. 2013) package to compute the overlap between these windows and TSS and TES of dosage compensated genes (considering on both strands). Then we computed the distance between the TSS (or TES) and the mid-point of the domain boundary.

Computing Insulation Score

The insulation score as defined in (Crane et al. 2015) is calculated on our data as the mean Hi-C signal in a 35Kb (10 bins) squared sliding window. We started from our 3.5Kb Hi-C matrices and computed the insulation score moving the squared sliding window along the main diagonal. We ignored the first and last 10 bins of the chromosome. We removed the non-informative diagonals: first two diagonals, i.e. interactions occurring at distance 0 and 1.

The insulation score values were then normalized by the mean insulation score of each chromosome as in the original study (Crane et al. 2015). Since the domain boundaries are defined at the intersection between the TAD start and end bins, the mean normalized insulation score from the two adjacent bins is considered.

Distribution of normalized CLAMP signal files

CLAMP ChIP-seq enrichment signal files (.wig files from GSE39271) were rescaled by dividing the signals in each file by their 99^{th} percentile. This conservative normalization was applied to facilitate comparisons across samples accounting for potential differences in ChIP efficiency.

The CLAMP binding sites were assigned to the 3.5Kb genomic bin overlapping the mid-point of the binding site itself. For each bin containing a CLAMP binding site, the highest wig signal was obtained within the bin $Start$ and End positions. This

signal value was allocated as the probable CLAMP summit within that bin. We then fetched the unique list of nearest CLAMP bin for each of the disappearing boundaries in the core-set and report the summit values for those bins.

4C tag enrichment near domain boundaries

The 4C data by Ramírez *et al.*(Ramírez et al. 2015) based on 18 probes were processed as is.

The 4C data by Schauer *et al.*(Schauer et al. 2017) based on 11 probes were instead further filtered as we noted larger differences between replicates for some probes. Namely, we discarded 4C data originating form a specific probe if the two replicates have ≥ 2 fold difference in the total number of sequenced reads. To further avoid unbalanced comparisons, for each pair of samples compared (e.g. S2 WT vs MSL2-i) we considered a specific probe only if the it has ≤ 1.5 fold difference in the total number of sequenced reads across the compared samples. Thus, we obtained a total of 76 high quality 4C-seq dataset across 11 probes.

We used a similar strategy as Ramirez *et al.* (Ramírez et al. 2015). First, we reassigned 4C read counts to our reference DpnII fragment ends table as obtained from the cooler package. Read counts per fragment were binarised thus assigning value of 1 to fragments with one or more overlapping reads, and a value of 0 to fragments without any overlapping read. Replicates are then further merged and converted to 1 or 0 values based on if a replicate contained any counts in the corresponding 4C-seq dataset.

To compute the 4C enrichment value (E), the fragments ($frag$) were further converted to their corresponding mid points positions (m) and a small 20Kb window (w_{small}) was extended on both sides aggregating (by summing) all values (v) within

that window. This sum was further divided by the sum of all values aggregated within a larger 600Kb window (w_{big}) used to estimate the expected background signal.

$$E_m = log_{10}\left(\frac{\sum_{i=m-w_{small}}^{i=m+w_{small}} v_i}{\sum_{i=m-w_{big}}^{i=m+w_{big}} v_i} + 1\right)$$

With $w_{big} \leq m \leq \left(l_{chr} - w_{big}\right)$ to avoid windows extending beyond the chromosome start or end. The enrichment value E constitutes the observed over expected 4C signal ratio and was $\log_{10}$ transformed with the addition of a pseudocount value of 1 for downstream analyses.

To summarize the average 4C enrichment signal around domain boundaries mid points (m_d), grouped by class, the 4C data associated to fragments are mapped to the corresponding Hi-C bin (b) and their mean enrichment value assigned to the bin (E_b). A window up to distance w from the boundary (m_d) is considered. The bins are then converted to their relative position (p) with respect to the bin containing the domain boundary mid (b_{md}). Thus, the enrichment values (E_b) are eventually assigned to their corresponding position (E_p) relative to any domain boundary ($\pm$35Kb). Finally, we compute the mean of enrichment values E_p for each position (p).

4.0 Results

4.1 About the Hi-C datasets presented in this study

The Hi-C data used for this project was originally generated using the simplified Hi-C protocol. This means that the restriction enzyme digested ends were not filled in using biotin tagged bases and the enrichment of ligation junctions containing biotin was not done (Sexton et al. 2012). The Hi-C data was provided by G Cavalli. We obtained on average 1 billion reads for each of the sex-sorted male and female embryo datasets. The male sample had 7 runs across two biological replicates and the female sample had 6 runs across two biological replicates (**Table 1**).

Table 1 Adapted from Pal et al., 2018 (manuscript in revision): Read statistics in sex-sorted male female *drosophila* Hi-C datasets

Sample	Replicate	Total Pairs	Discarded Pairs	Kept Pairs
Male	A1	160,609,800	146,322,053	14,287,747
Male	A2	158,119,428	144,025,206	14,094,222
Male	A3	160,559,337	146,270,219	14,289,118
Male	A4	160,178,817	145,910,435	14,268,382
Male	B1	177,854,486	161,473,733	16,380,753
Male	B2	176,891,839	160,551,750	16,340,089
Male	B3	180,632,426	164,470,001	16,162,425
Male	Merge	1,174,846,133	1,069,023,397	**105,822,736**
Female	A1	175,074,039	153,986,896	21,087,143
Female	A2	175,040,306	153,974,981	21,065,325
Female	A3	174,952,576	153,885,812	21,066,764
Female	B1	190,498,730	157,274,570	33,224,160
Female	B2	189,735,995	156,545,059	33,190,936
Female	B3	193,144,206	159,953,270	33,190,936
Female	Merge	1,098,445,852	935,620,588	**162,825,264**

Furthermore, we also sourced publicly available Hi-C data on S2 (male), Clone8 (male) (Ramírez et al. 2015) and Kc167 (female) (Li et al. 2015) cell lines. All Hi-C datasets were processed using the HiCLib (Imakaev et al. 2012). Additionally, we also obtained data generated using *insitu* Hi-C on *drosophila* S2 male cell lines from the same group for ongoing collaborations (Ogiyama et al. 2018). It is worth noting the effect these two techniques have on the filtering of Hi-C data. We employed the default filters used within the library and in addition we filtered for the sum of distances between mate pair mapping site and its nearest downstream restriction site. One of the key differences to note are the number of read pairs lost during the filtering of dangling ends (**Figure 3**). These are read pairs originating from un-ligated ends, or read pairs that are too close to each other. For the simplified Hi-C protocol on sex-sorted embryos, nearly one third of all read pairs were lost after applying this filtering step, whilst single-sided or read pairs where only one mate mapped to the genome was comparatively much lower. This would be expected from simplified Hi-C since the biotin enrichment was not done to enrich only those products which contained a ligation junction. Comparatively, for the *insitu* Hi-C protocol, dangling ends represent a very small proportion of read pairs filtered, whereas the single-sided read pairs represent a much larger fraction. Similarly, the cell lines, which followed the originally described Hi-C protocol show much smaller proportion of dangling ends.

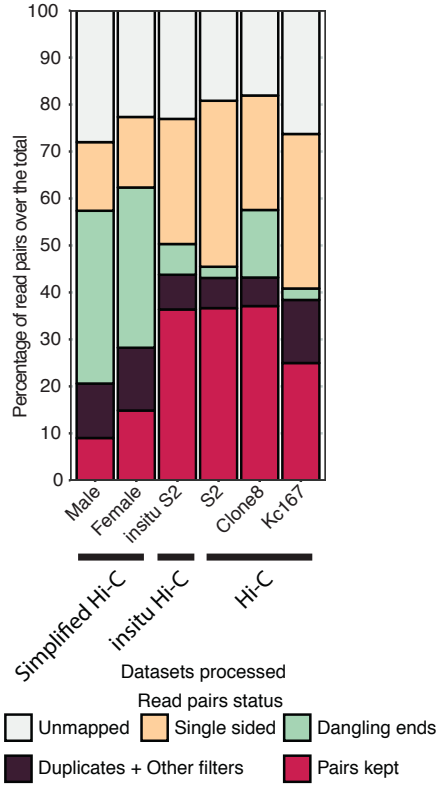

Figure 3 - **Read filtering statistics for the different Hi-C datasets**. All datasets in this study were processed with HiCLib (Imakaev et al. 2012). We processed *drosophila* embryo datasets generated using the Simplified Hi-C protocol. We also processed data generated using normal Hi-C for cell lines (S2, Kc167, Clone8). We also processed *insitu* Hi-C data in S2 for an ongoing collaboration. It is clearly evident, that not using the biotin enrichment step results in the generation of many more non-informative reads (read pairs resulting due to the sequencing of un-ligated fragments or those that are too near). On the other hand, protocols which make use of biotin enrichment show a much smaller fraction of dangling ends as these protocols all make use of the biotin enrichment step.

4.2 The number of reads left after processing varies between different pipelines

Although some Hi-C analysis pipelines have emerged as being highly popular today (Imakaev et al. 2012; Durand et al. 2016), at the inception of this project that was not the case. A significant amount of time was divested towards solving dependencies and testing different Hi-C analysis pipelines. Two different Hi-C processing workflows were used: HiCPipe(Yaffe and Tanay 2011) and HiCLib (Imakaev et al. 2012). HiCLib allowed for better modulation of the filters applied post-alignment. Furthermore, we were able to incorporate it into an easy-to-deploy pipeline built using BASH. For HiCPipe we used the default filters used by the entire pipeline. For HiCLib, post-filtering we obtained more than 100 million read pairs for both samples. Whereas, for HiCPipe we obtained 70 million for the male sample

and 124 million for the female sample (**Table 1**).

Furthermore, HiCPipe employs a very aggressive duplicates removal procedure by imposing an equivalency between all interacting fragments at the read pairs level by binarising all read counts. Therefore, even if two fragments have 100 read pairs validating them, in the HiCPipe workflow these two fragments are treated as having one read pair. Although this assumption may be sufficient at lower resolution Hi-C data, this may hinder downstream data analysis for Hi-C data binned at much higher coverage (Jin et al. 2013). Therefore, we report our analysis using data processed with HiCLib, whereas the usage of HiCPipe processed data is presented solely as an alternative analysis of interaction decays. The embryo datasets processed with HiCLib (Imakaev et al. 2012) were subjected to two different normalizations from the same library. The first, chromosome-wise ICE normalises each chromosomal Hi-C map without taking into account the *trans* or between chromosome contacts. Genome-wide ICE on the other hand takes into account these *trans* contacts, normalising each Hi-C map to the genome-wide average.

4.3 The male chromosome X Hi-C maps shows higher long-range contacts

The male X chromosome exists in a single copy state. Whereas, the female X chromosome exists in two copies. Assuming equal sequencing depth of the experiment, the male chrX is expected to contain at most half as many read pairs as the female chrX and the autosomes. The male chrX *cis* Hi-C map has on average 3.5 times less number of reads than any female cis Hi-C map and 2.2 times less number of reads than the male autosome Hi-C maps. Thus, we asked if there are any observable differences in the global interaction pattern between the male and female chrX. A log2 ratio was computed between the independently normalised (chromosome-wise ICE) chrX Hi-C maps. We expected that the male chrX should

display lower contact frequencies than the female, as the chrX exists in lower copy number. Yet, we noticed an increase in contact probabilities in the male Hi-C maps at longer distances (500Kb - 1Mb). At shorter distances, the female Hi-C map consistently has higher signal. Beyond these distances, the male chrX has higher signal and these interactions border regions in the Hi-C map that may correspond to TADs. This suggests that the increase in contact probabilities is observed in contacts that would occur between TADs in the inter-TAD contact space (**Figure 4**).

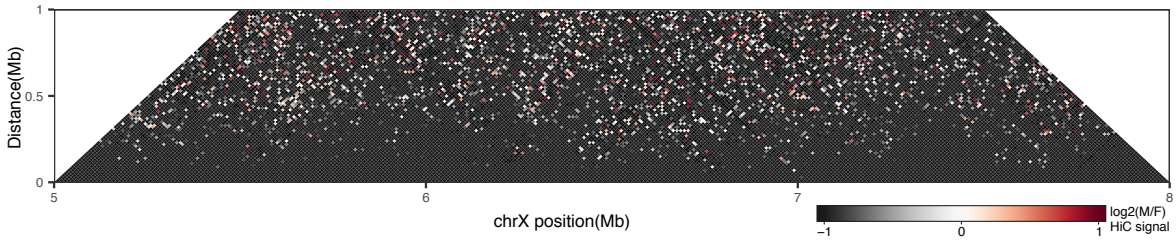

Figure 4 Adapted from Pal et al., 2018 The male X chromosome participates in more long-range contacts. Sex sorted embryo datasets binned at 10Kb were normalised chromosome by chromosome using HiCLib (Imakaev et al. 2012). Depicted above is a 3MB segment of chromosome X showing the fold change of Hi-C signal between male and female embryos. TAD structuring seems to be preserved for the most part (black triangular regions). In these regions the female Hi-C matrices consistently shows higher signal (black). The male single copy X chromosome starts to show more equivalent signal in the inter-TAD regions (dark blue regions) at distances greater than 250Kb. At distances greater than 500Kb, the male chrX consistently shows Hi-C signal that is nearly at par with the female chrX (dark blue) with interspersed regions showing very high contact frequencies (white).

4.4 Higher long-range contacts have a quantifiable effect on Hi-C signal decay

After this qualitative observation, I wanted to quantify and confirm these observations. Therefore, I computed the Hi-C interaction decay as a function of the distance for the sex-sorted male and female embryos, alongside additional male (Ramírez et al. 2015) and female (Li et al. 2015) cell lines. For this analysis, I considered each autosome arm independently of the other and removed chromosomes 4 and Y. We did this, owing to their smaller footprint, chr4 is approximately 6Mb in size, while the Y chromosome is 347Kb in size. Furthermore, the Y chromosome is mostly un-mappable due to the presence of repetitive regions (Charlesworth 2001) and chromosome 4 is mostly heterochromatic (Sun et al. 2000). Also, the Y chromosome itself is not directly linked to sex determination in drosophila (Samata and Akhtar 2018).

Log-log interaction decay plots were computed for each of the independently normalised chromosomal arms and the X chromosome. I noticed that in the log-log plot the chrX showcased a switching pattern between the male and female samples. This was observed in both the cell lines and the sex-sorted embryo datasets. Namely, as dosage compensation came into effect, the male chrX behaves differently from the autosomes (**Figure 5**). While all of the autosomes, except for chr2R behave in a similar fashion the chrX interaction decay slowly moved away from the autosomes and ended with a larger median interaction value at larger distances in the male sample. This is a result, confirmed using the sex-sorted embryos and two independently generated cell line Hi-C datasets (Ramírez et al. 2015; Li et al. 2015) from two different years. This shows the chrX having a much slower decay than the autosomes. Or in other words, the chrX had higher contact frequencies in the mid-/long-range distances.

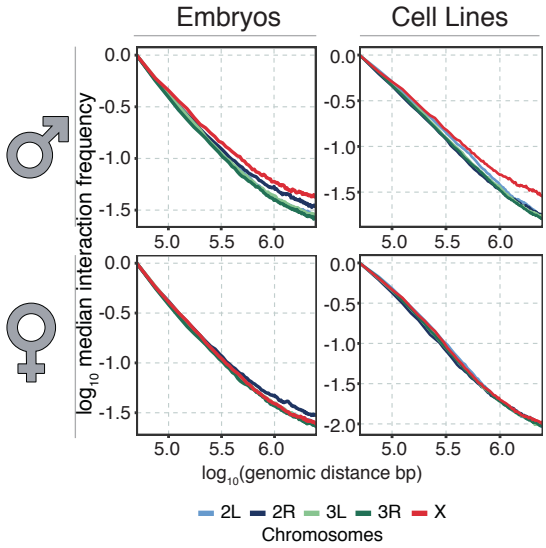

Figure 5 Adapted from Pal et al., 2018 Male chrX shows slower decay in Hi-C signal. 25KB binned Hi-C data for embryos and cell lines were normalised chromosome by chromosome using HiCLib (Imakaev et al. 2012). Shown, is the log10 median Hi-C signal (y-axis) at each genomic distance within the distance range of 50Kb to 2.5Mb (x-axis). The median Hi-C signal at each distance was additionally normalised by the Hi-C signal at the first distance considered (50Kb) to make the differences visually observable. Comparing the chrX between males and females across embryos and cell lines, we observed that although the decay starts at the same point (y-axis value 1), the chrX starts to move away from the autosomes at longer distances. This is observed in the male samples in both embryos and cell lines. But, this pattern is not present in either of the female samples.

To confirm that indeed the chrX has a slower decay rate, I computed the slope coefficients for each of the chromosomes. The differences (delta) between these slope coefficients (**Figure 6**) shows that the chrX slope is less negative than the autosomes in the male samples (average difference 0.11) and shows the rate of Hi-C signal decay is slower in the male chrX, which would be an effect observed due to more long-range contacts. Furthermore, we observed the same effect in both the sex-sorted embryos and cell lines datasets. This was not observed in either of the female samples. A similar behaviour observed in chr2R in both the sexes can be explained by the increased propensity to preferentially participate in trans interactions with chr2L as evidenced in **Figure 23**.

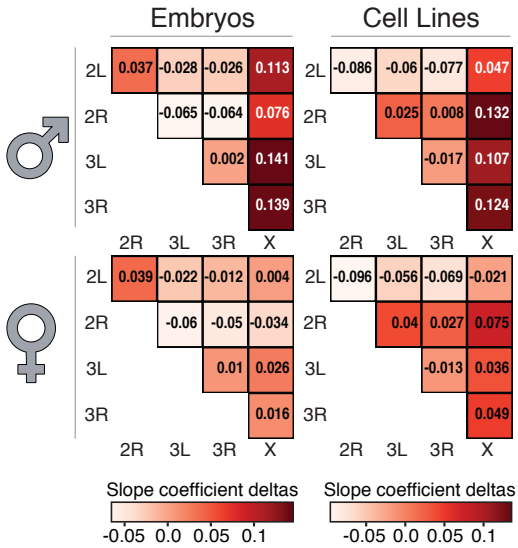

Figure 6 Adapted from Pal et al., 2018 The Hi-C signal decay is quantifiably slower in the male chrX. Using linear modelling between the distance ranges of 50Kb and 400Kb, the slope coefficients were estimated for the Hi-C signal decay profiles. The pairwise differences of slope coefficients between all chromosomes shows that the chrX has a less negative slope than the autosomes. This is observed in the male samples, but is not observed in the female samples.

4.5 The difference between the slope coefficients is not due to a difference in copy number

I reasoned that the difference in slope coefficients might be due to a difference in copy number and thus the coverage of Hi-C map. The autosomes being in higher copy number are at a higher sequencing depth and coverage. This means that the total number of events sampled by the autosomes from the ensemble space is theoretically higher than chromosome X. Therefore, the pattern that was observed may be a different signal saturation achieved in the respective chromosomal Hi-C maps. Furthermore, a number of interactions can be observed in the Hi-C maps (**Figure 4**) wherein the male Hi-C maps have extremely high contact frequencies. Although, the functional relevance of these interactions were not subjected to further investigation, it has been noted previously that normalization procedures in consort with lower density of Hi-C maps may sometimes create spurious signal (Rao et al. 2014). Therefore, the presence of these interactions and the lower overall signal density could be a bias affecting the slower Hi-C signal observed in the male

chromosome X.

To validate the hypothesis that the lower copy number of chrX is not causally linked to the slower interaction decay observed in the male samples, we randomly downsampled the Hi-C counts in the autosomal *cis* data, such that the total interaction frequency was half as much as the original datasets. We were able to confirm that random downsampling of the autosomes did not change the difference in interaction decays that we originally observed (**Figure 7**).

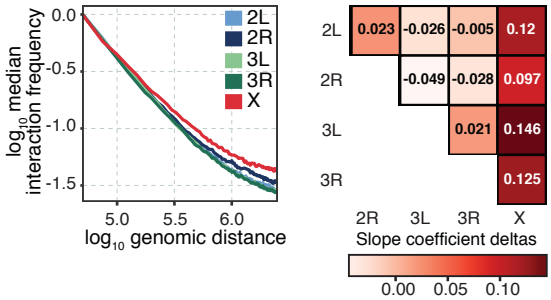

Figure 7 Adapted from Pal et al., 2018 The difference in the decay of Hi-C signal between chrX and autosomes is not due to differences in copy number. The 25Kb binned *cis* Hi-C matrices for the autosomes in males were downsampled randomly to simulate a condition where the autosomes have a single copy, such that the total signal originating from each autosome is half as much as the original Hi-C matrix. The log10 median Hi-C signal is plotted against the genomic distance ranging between 50Kb and 2.5Mb. The median Hi-C signal at each distance was normalised by the median Hi-C signal at the first diagonal (50Kb) (left). The slope coefficients were computed using linear modelling for values within the distance ranges of 50Kb and 400Kb and the pairwise comparison is shown (right).

4.6 The difference between the slope coefficients is not due to biases in biological replicates

Reproducibility of signal between replicates in Hi-C data has been widely investigated (Yardimci et al. 2017) and different methods have been proposed. Therefore, we were interested in investigating if the difference between slope coefficients was a bias from the different replicates used for the experiments. Different methodologies have been proposed for testing reproducibility have been proposed (Yardimci et al. 2017). We went a step further and tested both the reproducibility of Hi-C signal and effect across biological replicates. To investigate whether the differences in slope coefficients was an outcome of a bias in replicates we conducted similar analyses across both biological replicates in males and females. Hi-C contact matrices were independently normalised and compared across both biological replicates in males and females (**Figure 8**). The segregation pattern that we observed previously was still present in both replicates when computing the difference of slope coefficients between chrX and autosomes (**Figure 9**). Furthermore, the difference of slope coefficients are strongly correlated between both biological replicates in male and female Hi-C maps (**Figure 10**).

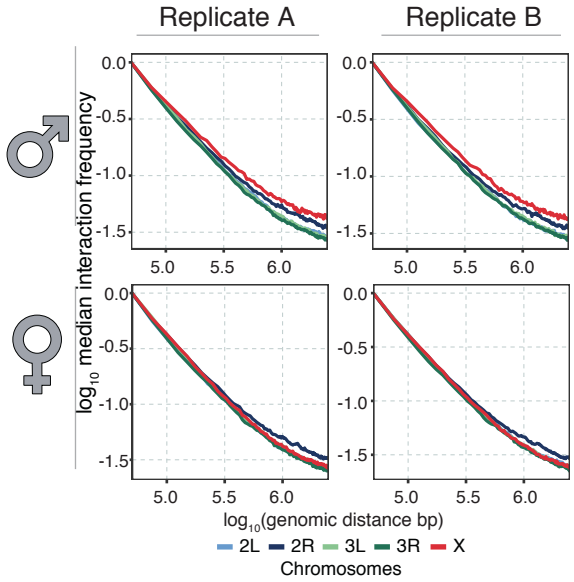

Figure 8 Adapted from Pal et al., 2018 The difference in the decay of Hi-C signal between chrX

and autosomes is not an effect due to biases in biological replicates. The 25Kb binned *cis* Hi-C matrices for each of the replicates in the male and female embryos were normalised chromosome by chromosome. For each chromosome the median log10 Hi-C signal is plotted against genomic distance in the range of 50Kb to 2.5Mb. The median Hi-C signal at each distance was normalised by the median Hi-C signal at the first diagonal (50Kb). The pattern wherein the chrX decay line moves away from the autosomes is still visible in the male samples, but is absent in the female samples.

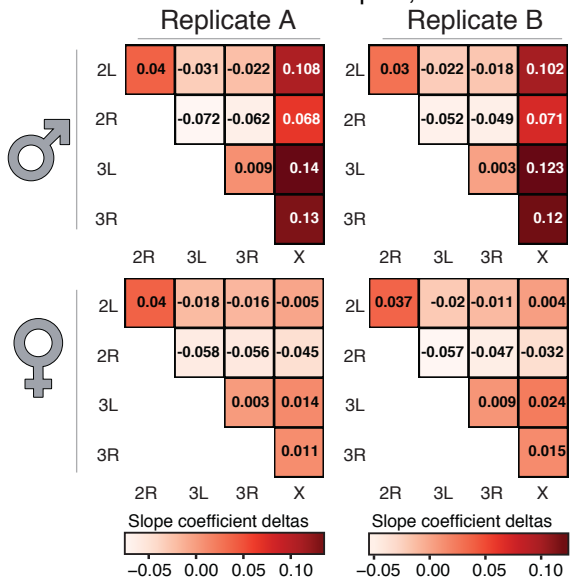

Figure 9 Adapted from Pal et al., 2018 The difference in the decay of Hi-C signal between chrX and autosomes is not an effect due to biases in biological replicates. The slope coefficients were computed using linear modelling for values within the distance ranges of 50Kb and 400Kb (**Figure** 8) and the pairwise comparison is shown for each of the replicates in the sex-sorted male and female embryos.

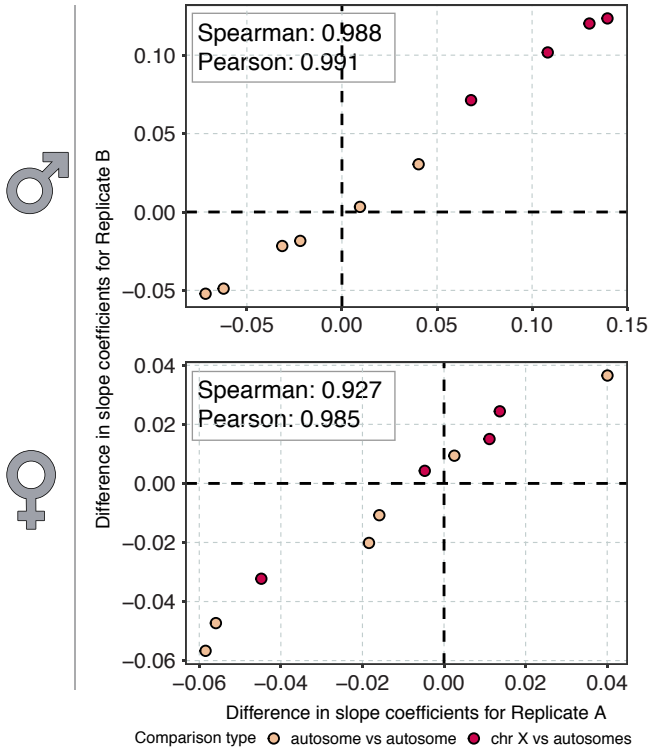

Figure 10 The difference in slope coefficients are highly correlated between the biological replicates. The pairwise differences of slope coefficients (**Figure** 9) are highly correlated between replicates in both male and female samples in the sex-sorted embryos and show very high spearman/pearson correlation.

To the reproducibility of Hi-C signal we also used a recently developed

method, HiCRep (Yang et al. 2017) to compute the correlation between both replicates in the male and female embryos. HiCRep smoothes the read counts in a Hi-C matrix and computes a stratum adjusted correlation coefficient (SCC) (Yang et al. 2017). Adjusting for the distances separating interacting genomic loci by assigning weights, HiCRep computes a correlation value between the two Hi-C matrices. The replicates show an average SCC value of 0.97 for the male embryos and 0.98 for the female embryos.

4.7 The difference between the slope coefficients is not due to the presence of extreme values in the Hi-C maps

One of the filters generally employed post-alignment is the filter for extreme values. Herein, the top 0.5% of contacts between genomic loci are removed and set to zero. This is done so that downstream statistics are not biased by the presence of these values. In analyses such as the ones depicted above, the analyses would be biased by extreme values only if the mean interaction frequency was used. We chose to keep these values and instead used the median interaction frequency, which is extremely robust to the tails of a distribution. Even though we never used the mean contact frequency in any of the analyses above, we can demonstrate that the presence of extreme values are neither biasing our observations or inferences.

We employed an extreme smoothing procedure by normalising all interaction values by the median interaction value at the starting diagonal. Any values which are greater than this median value are set to 1 (Giorgetti et al. 2014). Therefore, we are able to control for extreme values. Since Hi-C data follows a power-law equation, this transformation assumes that at distances greater than 20Kb, contact frequencies cannot be greater than the median contact frequency of genomic loci separated by a distance of 20Kb.

Using this procedure, we are able to show that the difference between chrX

interaction decay still exists between male and female embryos after smoothing extreme values (**Figure 11**). Furthermore, Hi-C normalisation methods are known to introduce spurious extreme values in sparse matrices (Rao et al. 2014). Using this same transformation across different normalization procedures we demonstrate that the difference between slope coefficients still exists after controlling for the extreme values which may have been present in the Hi-C experiment itself or may have been introduced by the normalisation procedure (**Figure 12**). We also note, that the hippie normalisation procedure binarises contacts between all genomic loci pre-normalisation. This makes all contacts equivalent. Even in this case, the difference is still present after binarization and smoothing of the distributions (**Figure 11 hicpipe**).

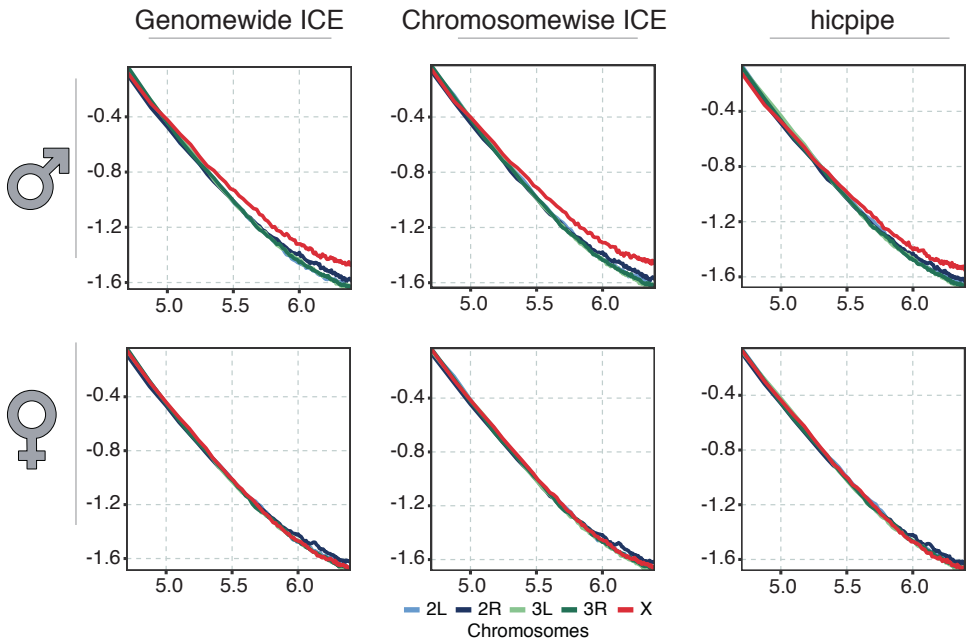

Figure 11 Adapted from Pal et al., 2018 The differences in Hi-C interaction decay are extremely robust to outliers and extreme values in the Hi-C maps. 25Kb binned Hi-C data for the sex-sorted fly embryos were normalised using three normalisation procedures. From left to right, chromosome by chromosome implicit normalisation using HiCLib (Imakaev et al. 2012), genome-wide implicit normalisation using HiCLib (Imakaev et al. 2012), and hicpipe (Yaffe and Tanay 2011) based explicit normalisation. The normalised interaction frequencies were then converted into probabilistic values by normalising with the mean interaction frequency at distance 50Kb (Giorgetti et al. 2014). Any values which were greater than this value was set to 1. It is observed, that even after this extreme transformation the pattern wherein the chrX decay line moves away from the autosomes is still visible in the male samples, but is absent in the female samples.

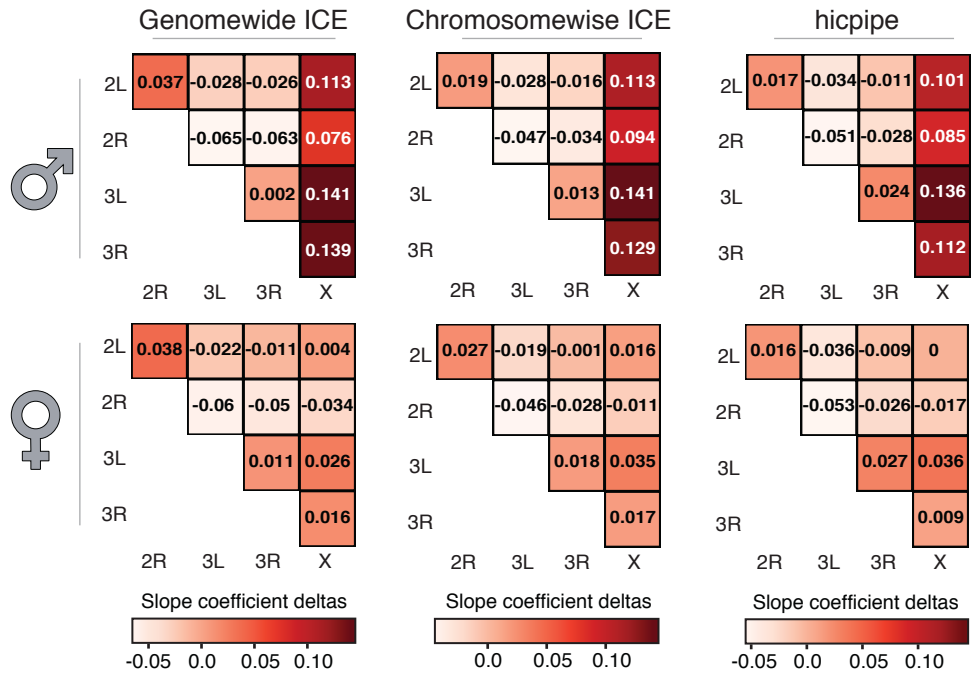

Figure 12 Adapted from Pal et al., 2018 The differences in slope coefficients are extremely robust to outliers and extreme values in the Hi-C maps. The slope coefficients were computed using linear modelling for values within the distance ranges of 50Kb and 400Kb (**Figure** 11) and the pairwise comparison is shown for the male (top) and female sex-sorted embryos across the three normalisation methods.

4.8 The difference between slope coefficients not an effect due to homologous pairing

We also considered a scenario wherein the pairing of homologous chromosomes may have an influence on the chrX specific differences in interaction decay. Homologous chromosomes are known to be paired throughout the cell cycle in *D. melanogaster*, although the exact molecular mechanisms are not completely characterized yet (Joyce et al. 2016). Between chromosome or *trans* interactions originating from the homologous chromosomes cannot be distinguished from the *cis* or within-chromosomal contacts. I reasoned, that since the male chrX is the only one without a pair, this may affect the interaction slope decay.

To understand what effects the homologous chromosomal pairing has on the interaction decay profiles, collaborators conducted polymer simulations on a paired and unpaired chromosome. Taking a single chromosome starting from a RabI-like configuration, molecular dynamics simulations were done using 10Kb beads on a

string in the presence or absence of preferential physical pairing across the entire length of the chromosome (**Figure 13**). With these simulations, we were able to show that in the absence of pairing, a faster decay is seen in the polymer simulation, which would translate into less long-range interactions. But, in our data, we observed an increase in long-range contacts resulting in a slower decay in the interaction decay analysis. Therefore, we were able to show that the difference in slope coefficients is not an effect due to lack of homologous pairing.

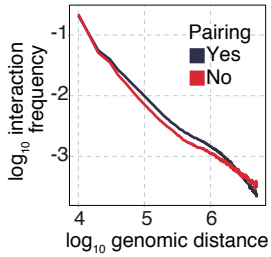

Figure 13 Adapted from Pal et al., 2018 Log-log plot shows the expected interaction frequency decay with distance obtained using molecular dynamics simulations. Male and Female sexes were considered with either a single or double copy of the X chromosome with or without the presence of pairing. Data shows the expected decay for a chromosome in double copy (with pairing) or single copy (no pairing) (black (Female) and red lines (Male), respectively). For any given binned genomic distance the log10 interaction frequency is reported (y-axis). Distances ranging from 10Kb to 5Mb are shown.

4.9 The difference between the slope coefficients is significant

Using non-parametric measures, we confirmed the significance of this difference in the rate of decay to be significant in both the male embryos and cell line (**Figure 14**). In other words, the deltas between chrX and autosomes is significantly different from the deltas between the autosomes. This pattern was not observed for the female samples. I also compared the normalised Hi-C signal in the male and female samples across implicit and explicit normalisation methods by computing the Kuiper's statistic. In this case too, we found the signal to be significantly different between chrX and autosomes (**Figure 15**).

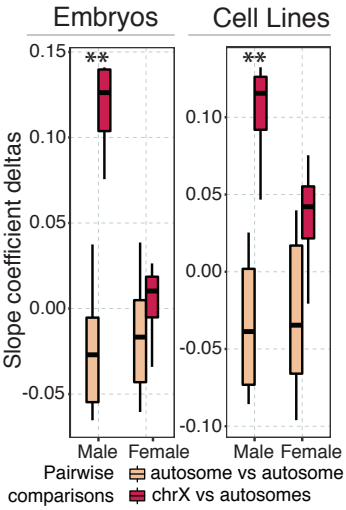

Figure 14 Adapted from Pal et al., 2018 The differences in slope coefficients are significant. Boxplot of slope coefficient differences for the Hi-C decay rates (**Figure 6**) grouped by autosomes or chrX in male and female embryos (left) or cell lines (right). The difference in rates of decay between autosomes and chrX chromosome is highly significant in male samples only (Wilcoxon test p-value 0.001).

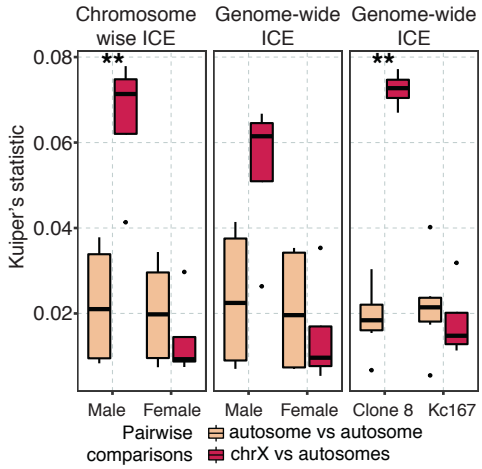

Figure 15 Adapted from Pal et al., 2018 Kuiper's statistic is reported as an alternative to comparing the the differences in slope coefficients. Kuiper's statistic is reported grouped by autosomes and chrX in male and female embryos (left and centre) or cell lines (right). We considered both chromosome by chromosome and genome-wide implicit normalisation using HiCLib (Imakaev et al. 2012). Kuiper's statistic is computed as the sum of absolute values for the maximum positive and negative differences between the cumulative density functions (CDFs) of the interaction probability for autosomes or chrX. CDFs of interaction probability were estimated from 50Kb to 2.5Mb as cumulative sums of median Hi-C contact frequencies for each distance, then divided by the cumulative sum maximum value to make it equal to probability 1.

4.10 A novel method to quantify structural differences between chromosomes using Hi-C data

We showed that the chrX participates in more long-range contacts. This is not an effect due to biological or technical biases. Yet, we were not sure if these long-range interactions were functional or random interactions occurring due to increased accessibility of the chromosome X. Functional interactions are those that occur between enhancers and promoters or between insulator binding regions. The identification of peaks or significantly interacting regions on Hi-C data using statistical methodologies is capable of identifying such interactions. One possible approach at our disposal was to identify such peaks in the male and female embryos. We could have then compared the proportion of dissimilar interactions that are between known/predicted regulatory regions. This would have allowed us to directly quantify the proportion of newly established long-range contacts in the male matrix that are between regulatory regions. It would have opened up the possibility for us to investigate the establishment of accessibility driven functional long-range interactions.

In our recent study (Forcato et al. 2017), we observed that all existing peak callers are extremely biased by a strong dependency between the number of peaks called and the coverage in the experiment. In our case, this was a problem as the single copy X chromosome *cis* Hi-C maps had half as many reads as the autosomes and three times less reads as the female X chromosome. Therefore, a direct comparison of peaks in the Hi-C data could not be made as the male and female Hi-C maps would never reach an equal statistical power. To resolve this issue, we devised a non-parametric approach that would allow us to compare the male and female Hi-C maps, irrespective of the difference in Hi-C signal between both samples.

In this procedure, we select the top-scoring interaction in any diagonal of a matrix. The selection of top-scoring interactions is based on setting a threshold on the quantiles. This threshold is set for each diagonal independent of every other diagonal. At the end of the selection, we are left with all interactions that were above the set threshold at that given diagonal. We call these interactions the "top-scoring" interactions (**Figure 16**). For example, using a threshold of 5%, we select the top 5% interactions in both the male and female Hi-C maps. Please note, that since we always select the top 5% at any given diagonal, we are left with the same number of top-scoring interactions in both datasets. What changes between the points is the spatial distance between adjacent top-scoring points located in different diagonals.

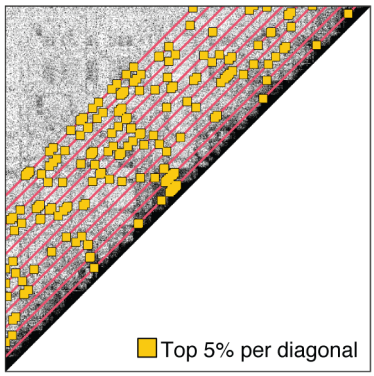

Figure 16 Adapted from Pal et al., 2018 An outline of the procedure for categorising top-scoring interactions are shown. Each diagonal represents a certain genomic distance separating interacting genomic loci. Progressively, interaction values are selected at each diagonal. Therefore, interaction frequencies are selected by the genomic distance separating the participating genomic loci. At each distance, the top 5% interactions are selected. These interactions are assigned a value of 1 and are called the top-scoring interactions (yellow).

I noted, that after completing this procedure on the male and female Hi-C maps, both Hi-C maps look remarkably similar on the autosomes, with most of the points being conserved across both male and female samples (**Figure 17**, left panel). This is not true for chrX, where most of the top-scoring interactions are conserved, but other regions can be seen which are also dissimilar and less clustered between the two (**Figure 17** right panel).

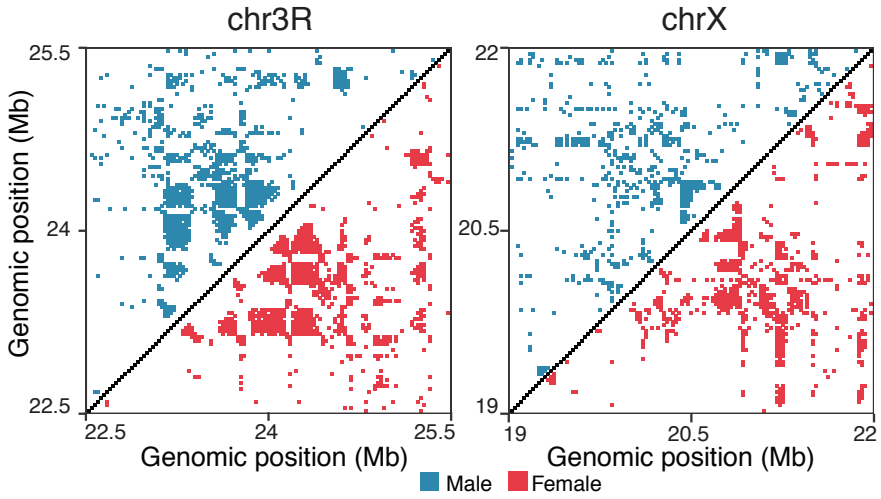

Figure 17 Adapted from Pal et al., 2018 Representative region of chr3R and chrX is shown after selecting top 5% interactions in the 25Kb binned male and female sex-sorted Hi-C datasets using the procedure outlined in **Figure** 18. Male is shown in the upper triangle (blue), whereas female is shown in the lower triangle (red). Since the top-scoring interactions are selected at each diagonal, the total number of top-scoring interactions are the same between male and female samples. But, the spatial positioning of these top-scoring interactions may not be same. Notice how similar both the male and female top-scoring patterns look in chr3R, but the same is not observed in the chrX.

We then devised a clustering procedure in euclidean space to quantify the difference in spatial distances between top-scoring interactions. My clustering procedure does not assume a pre-set number of clusters, but rather depends on the assumption that functional interactions will cluster with many other such interactions that may facilitate its occurrence. On the other hand, random interactions which are largely driven by accessibility will remain unclustered in space since there is no physical constraints to positively select such interactions.

4.11 The dosage compensated male chrX participates in more random interactions

Using the clustering procedure described above, I clustered the top-scoring interactions in the male and female Hi-C maps. I observed, that the differences between the clustered data points in female versus male is higher for chrX as compared to the autosomes. Namely, the male chrX shows more top-scoring interactions that could not be clustered (**Figure 18**).

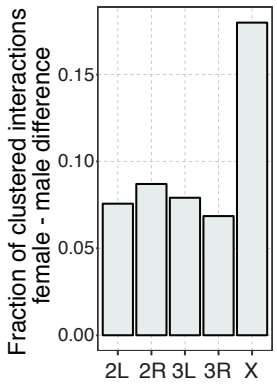

Figure 18 Adapted from Pal et al., 2018 Male chrX shows a higher proportion of unclustered top-scoring interaction. Top 5% interactions in the 25Kb binned male and female sex-sorted Hi-C datasets were clustered by iteratively aggregating all top-scoring interactions which were within a distance of 25Kb. This was done for all top-scoring interactions within a distance of 2.5Mb. Finally we report the difference in proportion of top-scoring interactions that could not be assigned/associated to a cluster/other nearby top-scoring data points.

We wanted to ascertain whether the observations were biased by different factors. First, I ascertained that the lower copy number of the male chrX could result in the generation of a larger fraction of randomly occurring top-scoring interactions. I also considered a scenario where the choice of normalisation method and parameters may affect which data points are categorised as top-scoring. Finally, I considered biases due to the experimental procedure which may be solely responsible for introducing more random top-scoring interactions in our analysis.

I demonstrate that even when considering different normalisation methods (chromosome-wise ICE, genome-wide ICE, hicpipe explicit biases modelling) and different percentile/distance thresholds for selecting and clustering the top-scoring interactions during the procedure, the chrX persistently clusters less (**Figure 19**). Furthermore, to check if the lower coverage of male chrX is generating more random top-scoring interactions, I randomly downsampled the female *cis* Hi-C maps, such that the total number of counts were equivalent to the corresponding chromosomal *cis* Hi-C map in the male embryos. My observations still hold true when comparing the downsampled female Hi-C matrix to the male Hi-C matrix (**Figure 19**).

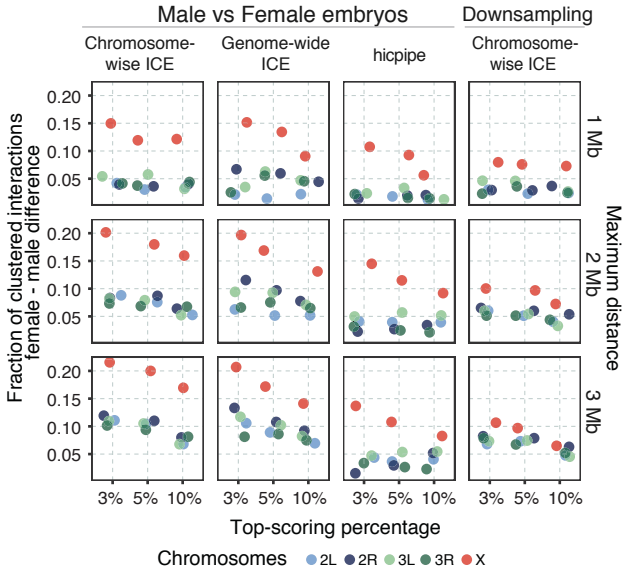

Figure 19 Adapted from Pal et al., 2018 The finding that chrX clusters less is not biased by algorithm parameters or copy number based differences in coverage. The 25Kb binned Hi-C matrices for the male and female sex-sorted embryos normalised with implicit (ICE) (Imakaev et al. 2012) or explicit (hicpipe) (Yaffe and Tanay 2011) procedures (top axis) were passed through different combinations of top-scoring interaction thresholds ranging from 3% to 10% (x-axis). Furthermore, the maximum distance separating sampled interactions was also iteratively changed in the ranges of 1MB to 3MB (right axis). To control for differences in copy-number the female chrX Hi-C matrix was randomly downsampled to the same coverage as the male chrX Hi-C matrix (right). In almost all cases, the difference between the proportion of unclustered top-scoring points between the male and female Hi-C datasets is higher for chrX than for autosomes.

Finally, I also used the clustering procedure on the three different *drosophila* cell lines; S2 (male), Kc167 (female), Clone8 (male). These male and female cell line Hi-C data were generated by independent groups, using independently standardised experimental protocols and machines. Even when comparing these datasets, I observed that the male chrX consistently clusters less than the female chrX (**Figure 20**).

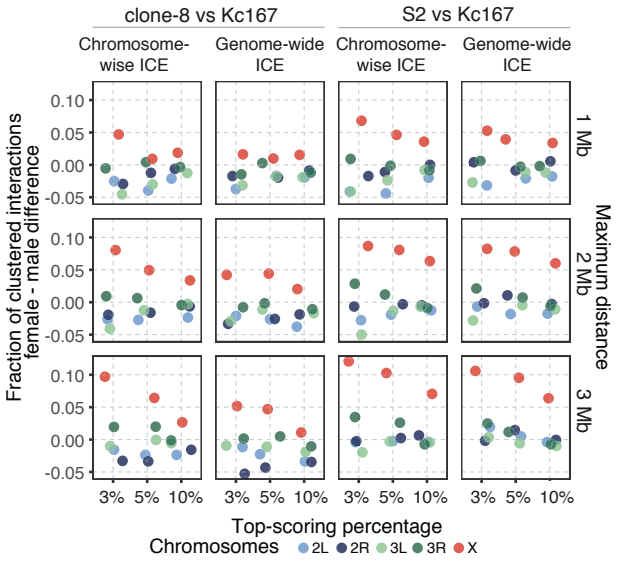

Figure 20 Adapted from Pal et al., 2018 The chrX also clusters less in male cell lines. 25Kb binned Hi-C matrices for the male and female cell lines (male Clone8, female Kc167, male S2) normalised with implicit procedures (Imakaev et al. 2012) were passed through different combinations of top-scoring interaction thresholds ranging from 3% to 10% (x-axis). The maximum distance separating sampled interactions was also iteratively changed in the ranges of 1MB to 3MB. In almost all cases, the difference between the proportion of unclustered top-scoring points between the male (Clone8, S2) and female (Kc167) Hi-C datasets is higher for chrX than for autosomes.

Taken in context with the previous results, even if the male chrX shows more mid-/long-range interactions compared to the female (**Figure 4**), the top-scoring interactions seem to be more randomly distributed (**Figure 17**). These results would be in line with a scenario wherein the dosage compensated male chrX is globally more accessible, thus more prone to participate in non-specific accessibility driven random non-functional events which can be detected in Hi-C data.

4.12 The dosage compensated male chrX is more accessible

I wanted to confirm that the random distribution of top-scoring interactions were indeed driven by increased chrX accessibility. To confirm increased accessibility, I investigated the inter-chromosomal Hi-C contacts as an estimation for non-specific interactions. Herein, the working hypothesis is that increased accessibility may result in more *trans* interactions. *Trans* interactions are those where either end of a read pair maps to different chromosome. I observed, that the chrX participates in more *trans* interactions. This has been observed across both embryos and cell lines (**Figure 21**). Therefore, the chrX is indeed more accessible.

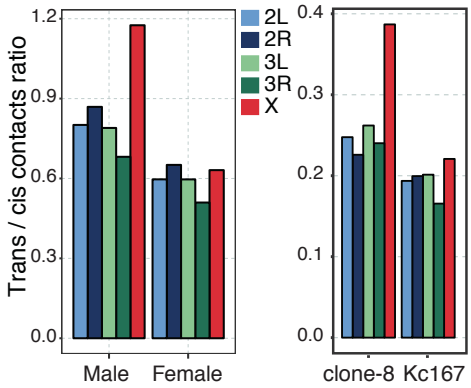

Figure 21 Adapted from Pal et al., 2018 The male chrX participates in more *trans* interactions than it does in *cis* interactions. The ratio of the total number of *trans* read pairs over *cis* read pairs is shown for each chromosome in male and female Hi-C datasets for both embryos and cell lines.

Furthermore, I considered a scenario wherein this effect might be due to the non-existence of pairing in the male chrX. The interactions occurring between homologous chromosomes are generally captured as *cis* interactions. But the homologous pairing effect is not present in male *drosophila*. The absence of this might lead to the effect wherein a relatively higher *trans* interaction is observed in the male sample. Thus, the effect of homologous chromosome pairing may introduce an additional bias factor. I tested the effects of copy number on the *trans* distribution and verified that the sex-sorted embryos *trans* interactions from the autosomes are specifically enriched on the male chrX (**Figure 22**). Yet, the same

effect is not observed in the female samples. Also, we were not able to confirm these findings in the cell lines (**Figure 23**).

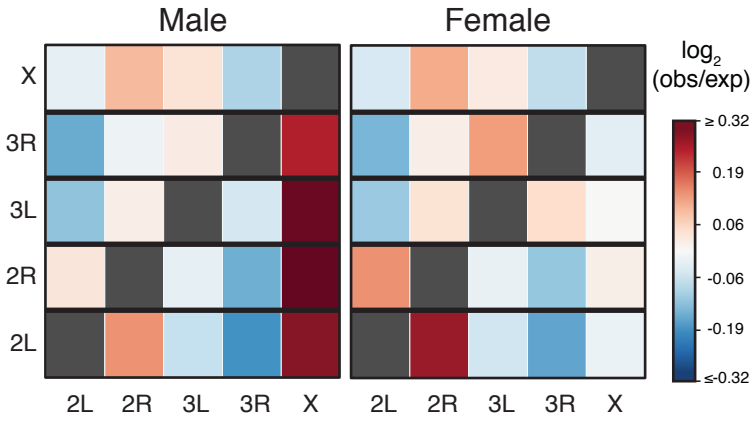

Figure 22 Adapted from Pal et al., 2018 The male chrX has a higher propensity to participate in *trans* interactions when compared to autosomes or the female dataset. The trans interactions for each chromosome (rows) was divided by a random expected value. Assuming a uniform distribution, the random expected value estimates the expected fraction of *trans* interactions belonging to the chromosome (rows) from the total *trans* interactions from any partner chromosome (columns) after adjusting for copy number (Methods). The log2 ratio of observed over expected fraction of *trans* interactions is reported in the heat map. This heatmap is not symmetric as the expected number of interactions is different depending on the origin vs target chromosome pairs. The diagonal is grey as cis-interactions are not considered.

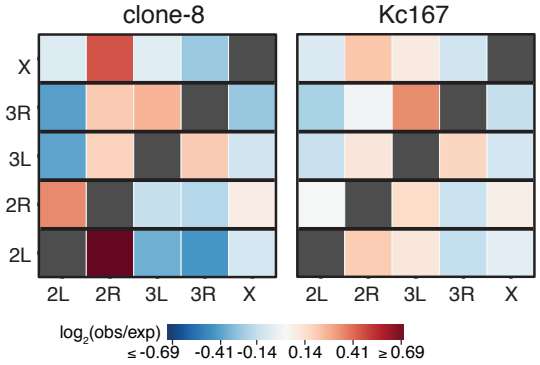

Figure 23 Adapted from Pal et al., 2018 The propensity of chrX to participate in more *trans* interactions is not observed in the cell lines. The trans interactions for each chromosome (rows) was divided by a random expected value. Assuming a uniform distribution, the random expected value estimates the expected fraction of *trans* interactions belonging to the chromosome (rows) from the total *trans* interactions from any partner chromosome (columns) after adjusting for copy number (Methods). The log2 ratio of observed over expected fraction of *trans* interactions is reported in the heat map. This heatmap is not symmetric as the expected number of interactions is different depending on the origin vs target chromosome pairs. The diagonal is grey as cis-interactions are not considered.

Therefore, I conclude that the dosage compensated chrX is more accessible and thus more prone to make random long-range interactions detectable using Hi-C.

4.13 A novel method for detecting genome compartmentalisation

Eukaryotic genomes have been progressively compartmentalised into higher order ensemble folding structures. Starting from the largest to the smallest, the genome has been compartmentalised into large-scale compartments (Lieberman-Aiden et al. 2009) that correlate with previously described band domains on the basis of trypsin digestion susceptibility (Manuelidis 1990) or on the basis of GC content (Saccone et al. 1993; Bernardi 1995). Compartments correlate with active and inactive regions of the genome (Lieberman-Aiden et al. 2009) and change between differentiation states (Dixon et al. 2015). With increases in Hi-C sequencing depth, compartments have been further compartmentalised into folding structures popularly termed as Topologically Associated Domains (TADs) (Dixon et al. 2012; Sexton et al. 2012). TADs are regions of aggregated chromatin in Hi-C maps, wherein regions of chromatin that are distant in linear space tend to contact each other more than their adjacent neighbour.

TADs are largely invariant and do not change between differentiation states (Dixon et al. 2015). Rather, the insulation between TADs change between differentiation states (Dixon et al. 2015). Insulation, is a metric that quantifies the separation between TADs as a ratio of the intra-TAD versus inter-TAD contact frequency (Crane et al. 2015; Zhan et al. 2017). TADs are bounded by directionally oriented insulator proteins such as CTCF (Shih and Krangel 2013; Rao et al. 2014) or BEAF32/CP190 in *drosophila* (Sexton et al. 2012). Although, the removal of these proteins is not enough to drive large scale changes in TADs, resulting in changes in local insulation (Nora et al. 2017; Shih and Krangel 2013). In eukaryotes, TADs are correlated with transcriptional states (Rowley et al. 2017) and early/late replicating regions of the genome (Pope et al. 2014). In drosophila, the appearance of TADs coincide with activation of transcription during zygotic development (Hug et al.

2017).

I was therefore motivated to see if dosage compensation in *drosophila* resulted in any detectable changes in TADs. The dosage compensation mechanism in drosophila specifically targets active genes on the X chromosome and up regulates them by a non-constant factor. This up-regulation results in an average genome-wide up-regulation of two-fold. A multitude of previous studies have shown that TADs are largely invariant and are highly resilient to change (Dixon et al. 2015; Nora et al. 2017; Rodríguez-Carballo et al. 2017). Furthermore, a slew of literature suggests a causal link between transcription/replication and TADs (Pope et al. 2014; Rowley et al. 2017; Hug et al. 2017).

The identification of TADs has generally been based on detecting a change in a global distribution. These distributions quantify various metrics. Most popular are the insulation based methods (Crane et al. 2015; Zhan et al. 2017). Within TAD regions showcase a different interaction decay profile compared to the outside TAD regions (Fudenberg et al. 2016). Interaction decay profile based methods have also been proposed (Weinreb and Raphael 2015). One of the first metrics proposed, directionality index (Dixon et al. 2012), is a bias metric that quantifies the propensity of each genomic loci to participate in either upstream or downstream interactions in a Hi-C map. If two regions are very far away and have high contact frequency, this translates to highly positive (downstream) bias for one region and a highly negative (upstream) bias for its partner region. The regions where a highly negative bias changes to a highly positive bias are generally the regions that mark the boundaries of TADs (Dixon et al. 2012).

Based on literature evidence, I assumed that transcriptional activity is one of the primary factors required for TAD emergence (Hug et al. 2017; Rowley et al. 2017; Le Dily et al. 2014). Since different genes have highly variable expression

values, the effect of most genes would result in local perturbations on the chromatin fibre. These changes would not be detectable on a global distribution, but can be detectable when using the local distribution. On the basis of this hypothesis, I proposed Local score differentiator (LSD).

I use the directionality index (DI) to measure the upstream/downstream bias (**Figure 24** top). From this bias, I compute the difference of DI between neighbouring loci to quantify the consecutive changes in biases (**Figure 24** middle). At adjacent loci, where one has a highly negative DI and the other a highly positive DI, the difference between these two Di values will be highly positive or highly negative (**Figure 24** middle). Using these delta values, I detect the change points where a highly negative DI becomes highly positive as outliers in a local DI distribution (**Figure 24** bottom).

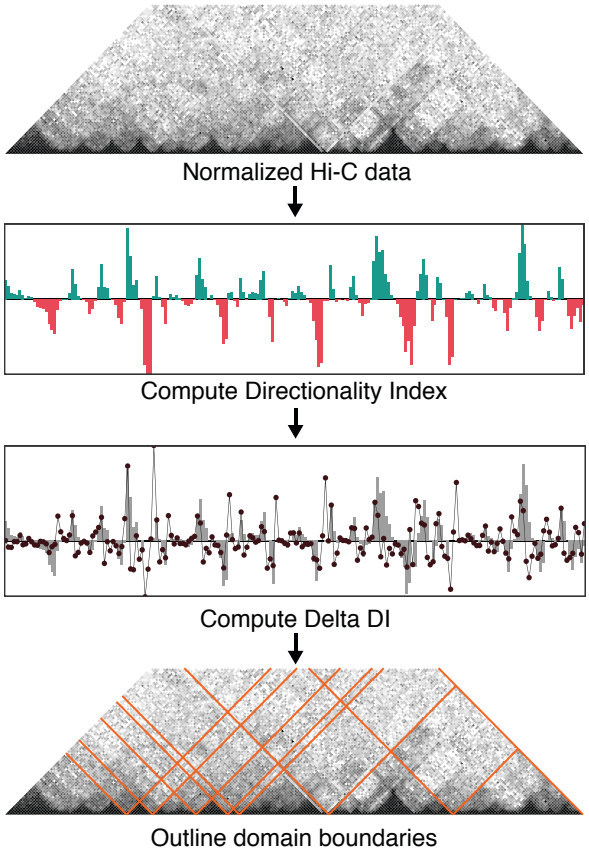

Figure 24 Adapted from Pal et al., 2018 Local Score differentiator (LSD) is a simplified approach towards detecting TAD boundaries. Starting from any normalised Hi-C matrix, the directionality index (DI) (Dixon et al. 2012) is first computed over a user defined window size. Next, the first derivative of the DI is computed (delta DIs). Then using a sliding window across the genome, local outliers are detected in the delta DI distribution as TAD boundaries.

4.14 Local score differentiator is extremely fast and accurate

In a recent study, we compared existing TAD calling procedures in terms of their true positive rate (TPR) and false discovery rate (FDR) when identifying TADs on simulated datasets (Forcato et al. 2017). I vetted LSD against these TAD calling procedures. In terms of calling TADs on these simulated Hi-C datasets, LSD outperformed all other TAD calling procedures in terms of true positive rate (TPR) and false discovery rate (FDR) (**Figure 25**).

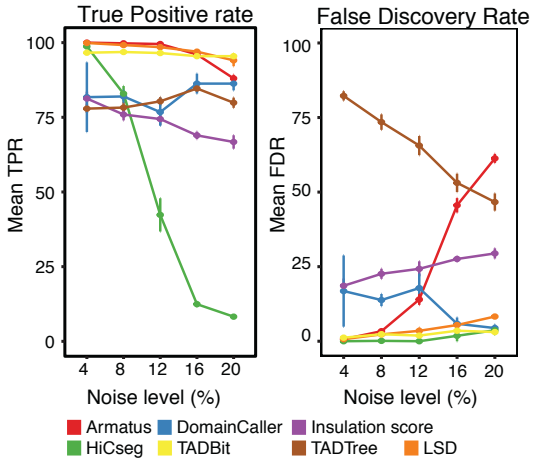

Figure 25 Adapted from Pal et al., 2018 LSD boundary calls are extremely robust in terms of true positive rate and false discovery rates. In a recent study we simulated Hi-C data (Forcato et al. 2017) and compared existing TAD calling algorithms. I used these simulated data to generate TAD boundary calls with LSD. In both cases, LSD showcased extremely high signal to noise ratio and showcased the highest TPR and lowest FDR. TADBit is the only other TAD calling algorithm, which performs as well as LSD.

TADBit (Serra et al. 2017) was identified as one of the most accurate TAD calling procedures in our previous study. Indeed, TADBit was also the only other TAD calling procedure that is similar to LSD in terms of performance. Therefore, we compared the speed of both TAD calling procedures on human 10Kb matrices from the Rao et al., 2014 study. TADBit is very slow, with the largest matrix belonging to chr6 taking nearly 10 days to complete (**Figure 26**). On the other hand, LSD processed the same matrix in less than 6 minutes. This represents an improvement of nearly 2500%.

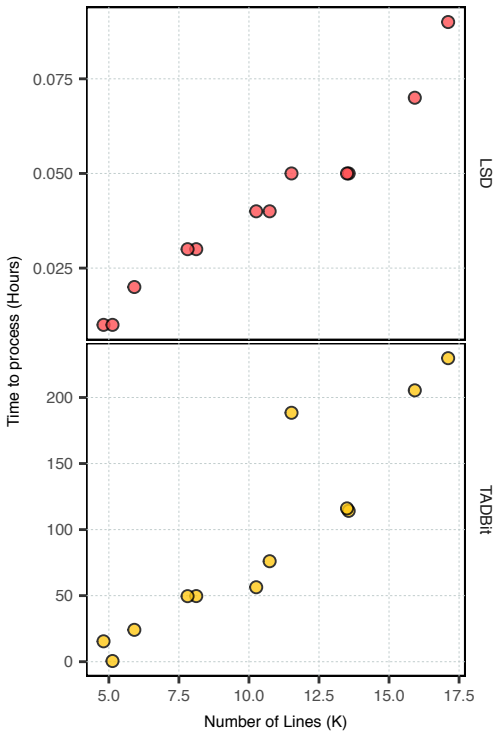

Figure 26 Adapted from Pal et al., 2018 Using 5Kb human Hi-C data from the Rao et al., 2014 study, I did TAD calls with both LSD and TADBit, to compare the speed of both these algorithms. LSD (top) is faster than TADBit (bottom) by a factor of nearly 2500%.

4.15 Chromosome X shows a higher proportion of non-matching TAD boundaries

I then applied this TAD calling procedure to the independently normalised male and female Hi-C datasets binned at 10Kb. Across various combinations of parameters for computing the DI and detecting outliers, I observed that the chrX consistently shows a higher fraction of non-matching TAD boundaries between the male and female embryos (**Figure 27**).

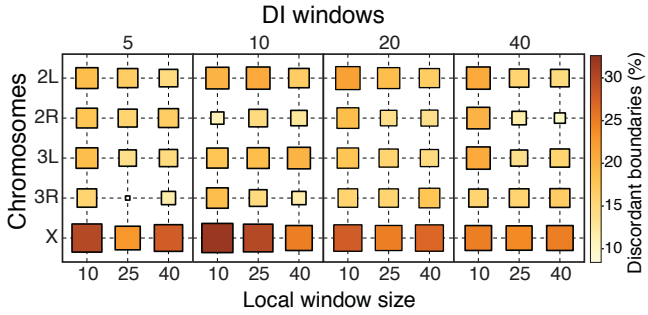

Figure 27 Adapted from Pal et al., 2018 The chrX consistently shows a higher fraction of non-matching TAD boundaries between male and female samples. Using various analysis parameter combinations LSD was used to do TAD calls on sex-sorted male female Hi-C datasets binned at 10Kb and normalised chromosome by chromosome with ICE. I iterated over two parameters, the size n of the directionality index window (upper axis) and the size m of the local window to scan for outliers (x-axis) across each chromosome (y-axis).

To ensure that my boundary calling procedure was not introducing technical biases in the analysis, I also used three additional TAD calling procedures (Filippova et al. 2014; Dixon et al. 2012; Serra et al. 2017). In all three cases, I observed a higher proportion of non-matching TAD boundaries in the chrX (**Figure 28**).

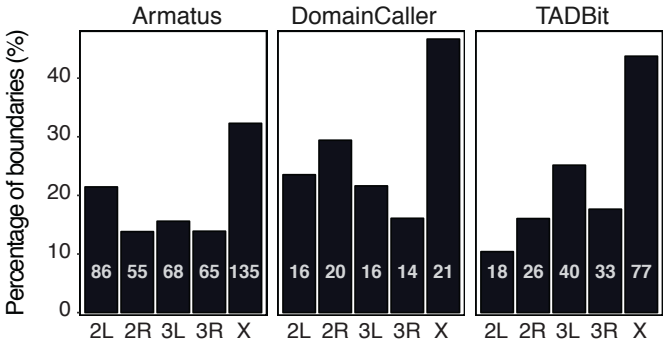

Figure 28 Adapted from Pal et al., 2018 The chrX also shows a higher fraction of non-matching TAD boundaries between male and female samples across other TAD calling approaches. Using default parameters, we did TAD calls using published TAD calling procedures (Filippova et al. 2014; Dixon et al. 2012; Serra et al. 2017) on sex-sorted embryos Hi-C data binned at a resolution of 10Kb and normalised with ICE chromosome by chromosome. Shown, are the proportion of non-matching boundaries between male and female samples using Domaincaller (Dixon et al. 2012), armature (Filippova et al. 2014) and TADBit (Serra et al. 2017).

I then tested if my observations were biased due to various factors such as the binning resolution, choice of normalisation and the copy number difference in chrX between males and females. To control for copy number, I downsampled the female chrX and predicted TADs using LSD. Across various parameter scales I observed that the chrX consistently showed a higher proportion of non-matching domain boundaries (**Figure 29**). Using LSD, I am also able to show that my

observations are not biased by the choice of normalisation, binning resolution or parameter setting during the TAD calls (**Figure 30**).

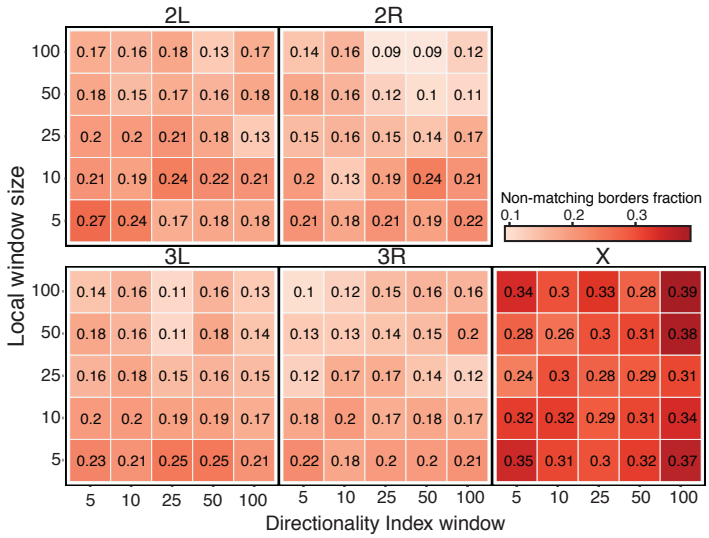

Figure 29 Adapted from Pal et al., 2018 The higher proportion of non-matching boundaries between male and female samples is not a function of the copy number difference. The 10Kb binned chrX Hi-C data for female embryos were downsampled to match the coverage of their corresponding male chromosome. The proportion of non-matching TAD boundaries between male and female samples is shown iterating over various combinations of DI windows (x-axis) and Local windows (y-axis). Colour intensity is mapped to their corresponding numbers.

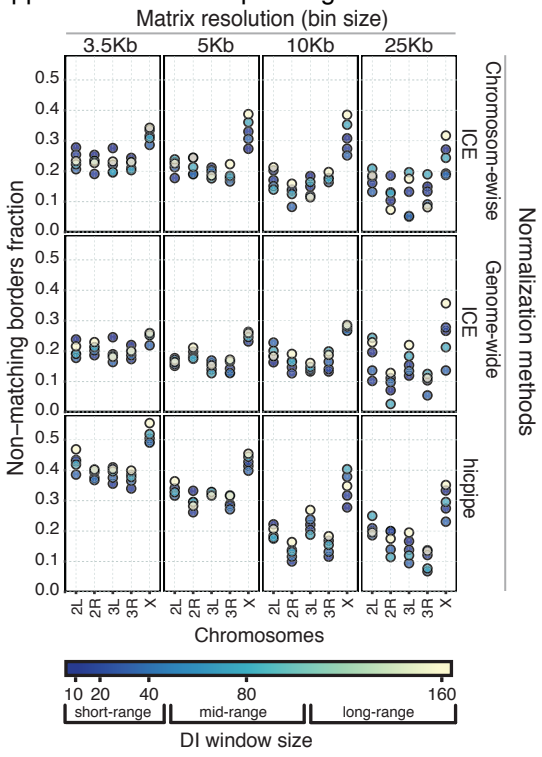

Figure 30 Adapted from Pal et al., 2018 The higher proportion of non-matching boundaries between male and female samples is not a function of normalisation or binning resolution. TAD calls using 5 different DI window sizes with a corresponding Local window size which is twice the size of the DI window, shows that the higher proportion of non matching domain boundaries (y-axis) in the chrX (x-axis) between male and female embryos is not linked to the binning resolution (upper axis) or the choice of normalisation (right-axis). The choice of the DI parameters were motivated by selecting parameters that would sample values from short-range, mid-range or long-range interactions.

4.16 Qualitative classification of non-matching domain boundaries correlates with dosage compensation

In order to verify if there was an association between dosage compensation and the non-matching domain boundaries. I binned the embryo datasets at 3.5Kb, the highest resolution possible for our data and using LSD, I predicted TADs and reduced the TADs to their respective boundaries. I then moved to assign a qualitative classification to these boundaries, so as to make a comparison between the male and female embryos. Boundaries which were found in both male and female embryos are labelled as Same, boundaries which were found only in the male sample were called as Appearing and boundaries found only in the female sample were called as Disappearing (**Figure 31**). In total, 851 boundaries were found on chrX across both male and female embryos. Of these 851 boundaries, 377 (44.3%) were categorised as same, 174 (20.4%) were categorised as appearing and 300 (35.3%) were categorised as disappearing boundaries.

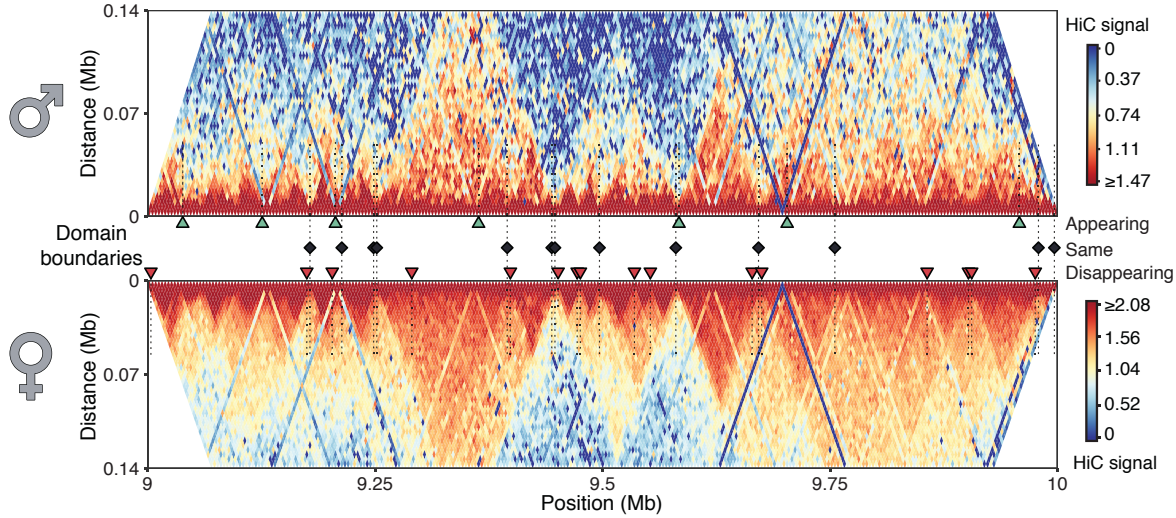

Figure 31 Adapted from Pal et al., 2018 The non-matching boundaries in chrX can be qualitatively categorised. TAD calls were done on 3.5Kb binned Hi-C matrices for the sex-sorted male and female embryos normalised chromosome by chromosome with ICE (Imakaev et al. 2012). We noticed, that at regions where boundaries weren't matching between male and female, specifically cases were observed where a boundary was identified in the female sample, but not in the male sample. In such cases, the Hi-C data also seemed to be a bit blurry and the separation between adjacent TADs was less clear. To annotate these changes, we created three categories. Same boundaries are those that were identified across both samples, Appearing boundaries are those that are identified only in the male sample and finally Disappearing boundaries are those that were identified only in the female sample.

We saw, that in many cases the qualitative assignment was able to explain our observations. A number of disappearing domain borders coincide with regions showing weakened insulation in the male sample, but the structures are still visible in the female sample. To quantify these differences, I computed the insulation score (Crane et al. 2015) for each boundary. The insulation score tries to quantify the number of interactions occurring across a boundary. The lower the insulation score, the less interactions occur across the genomic loci (Crane et al. 2015), as would be expected from domain boundaries. Comparing the insulation score at each disappearing boundary between male and female, we find that disappearing boundaries are significantly less insulated than the same region in female embryos (wilcoxon test p-value 0.001) (**Figure 32**).

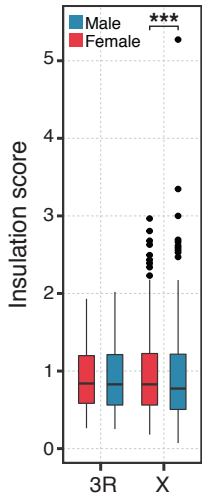

Figure 32 Adapted from Pal et al., 2018 Disappearing boundaries in chrX show higher change in insulation that those in autosomes. We adapted the insulation score metric as previously described (Crane et al. 2015) and compared the insulation of TAD boundaries between male and female samples. We observed, that chrX (right) disappearing boundaries showcased a significant change in insulation between male and female sex-sorted embryos (Wilcoxon test p-value 0.001) when compared to the autosome chr3R (left).

Furthermore, I found that MSL binding sites are enriched near the disappearing domain boundaries. This is true for all definitions of MSL binding sites (HAS or CES) generated by 3 independent groups. I also found that disappearing domain boundaries are closer to the TSS of dosage compensated genes (genes

which are up-regulated by the dosage compensation machinery) (**Figure 33**).

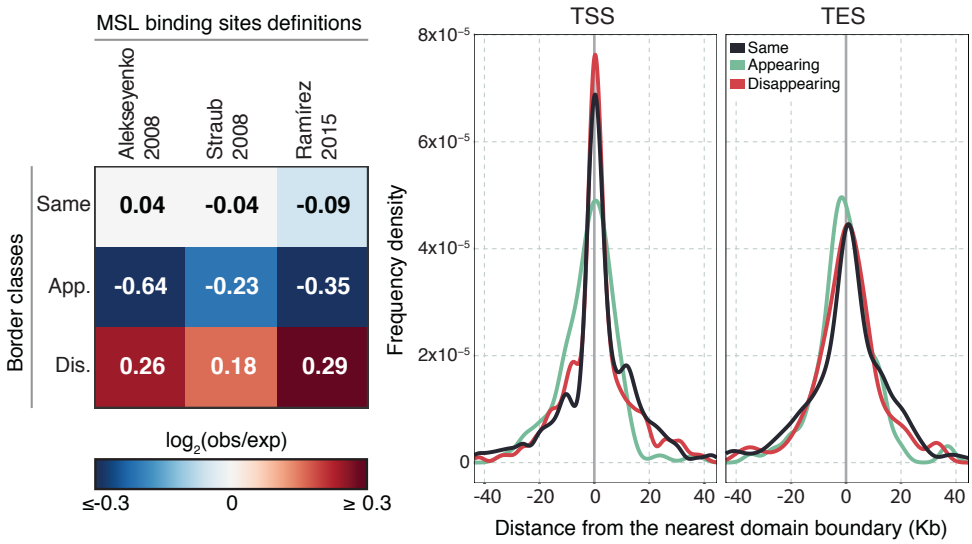

Figure 33 Adapted from Pal et al., 2018 Disappearing boundaries are associated to dosage compensation features. Left, log2 enrichment of dosage compensation complex binding sites is shown around domain borders. Three definitions of dosage compensation binding sites have been used from three different laboratories: Kuroda (Alekseyenko et al. 2008), Becker (Straub et al. 2008) and Akhtar (Ramírez et al. 2015) laboratories. The expected frequency was computed based on random uniform distribution of such sites along the chrX. Right, the frequency density of domain boundaries near dosage compensated gene TSS (left) or TES is shown.

4.17 Changes in insulation are not correlated to changes in insulator binding profiles

There are two affecters that may influence this change in insulation. The first, is a transcriptionally coupled change in insulator binding. This may translate towards a real change in higher-order chromatin structure. The second, relates to an increase in open chromatin regions near dosage compensated genes and dosage compensation complex binding sites. If more open chromatin regions are present near weakening boundaries, these boundaries will have a higher propensity to interact with their neighbours, resulting in higher contact frequency as compared to the female sample. To de-convolute these two possibilities, we first investigated the insulator binding landscape in S2 (male), Kc167 (female) cell lines alongside mixed embryos. We selected three principal insulator proteins; BEAF32, CP190 and CTCF from the modENCODE project (Contrino et al. 2012). Although, we saw a few specific changes, we did not observe a general pattern of association between insulator binding and the changes in insulation (**Figure 34**). Near equivalent proportion of domain boundaries are present near insulator peaks as defined by modENCODE (**Figure 35**) across both cell lines and embryos. The average number of insulator peaks overlapping domain boundaries is also very similar across cell lines and embryos (**Figure 36**). Finally, the average insulator binding intensity around domain boundaries is also very similar across both cell lines and embryos (**Figure 37**). In neither case did we observe an enrichment of CTCF profiles. This is in line with previous studies that have demonstrated that *drosophila* domain boundaries are not strongly associated to CTCF (Hou et al. 2012; Sexton et al. 2012).

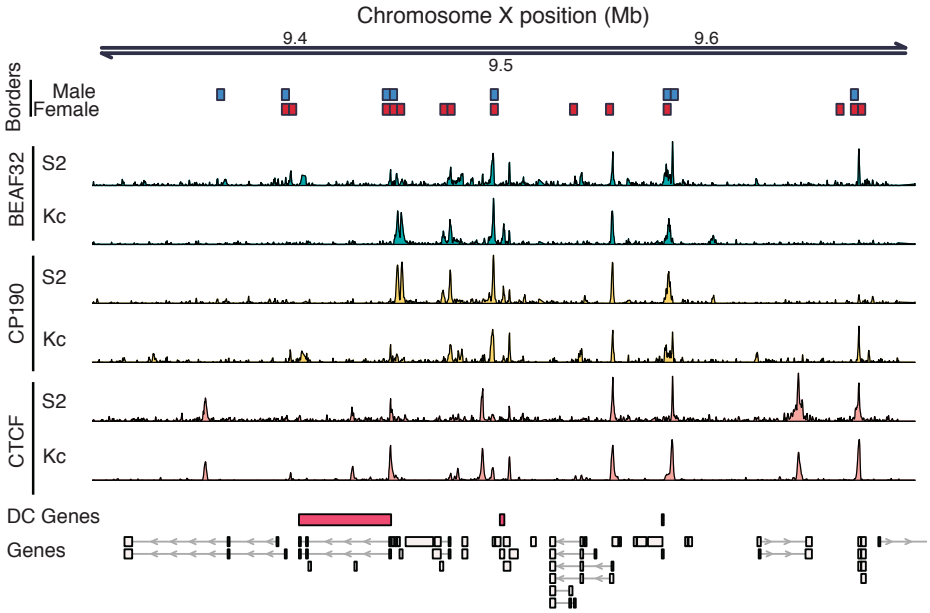

Figure 34 Adapted from Pal et al., 2018 Insulator profiles remain largely unchanged between different boundary classes and sexes. I obtained several publicly available insulator profiles from modENCODE (Contrino et al. 2012). Principally, I looked into profiles for BEAF32 and CP190, the major insulator proteins present in drosophila. Additionally, I also considered CTCF. To control for unknown biases, I tried to limit the search to insulator profiles which were produced by the same lab in the same year. Above, a small 500kb segment of chrX is shown with the various insulator profiles across S2 and Kc cell lines. Although, certain specific examples of insulator binding changes are evident, there does not seem to be any generalised differences in insulator binding.

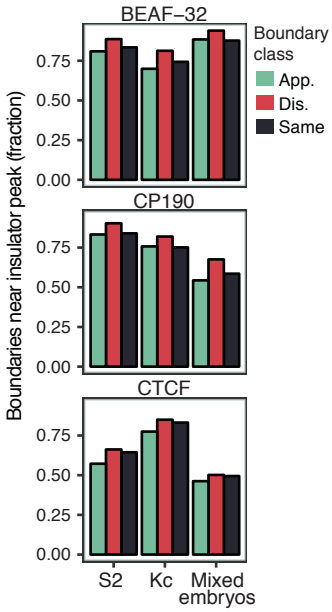

Figure 35 Adapted from Pal et al., Nat. 2018 Almost all domain boundaries, irrespective of their boundary change class are near an insulator peak as identified by modENCODE. The fraction of boundaries which are near (within 10 bins or 35Kb) of such an insulator peak are shown for BEAF32, CP190 and CTCF across S2, Kc and mixed embryos.

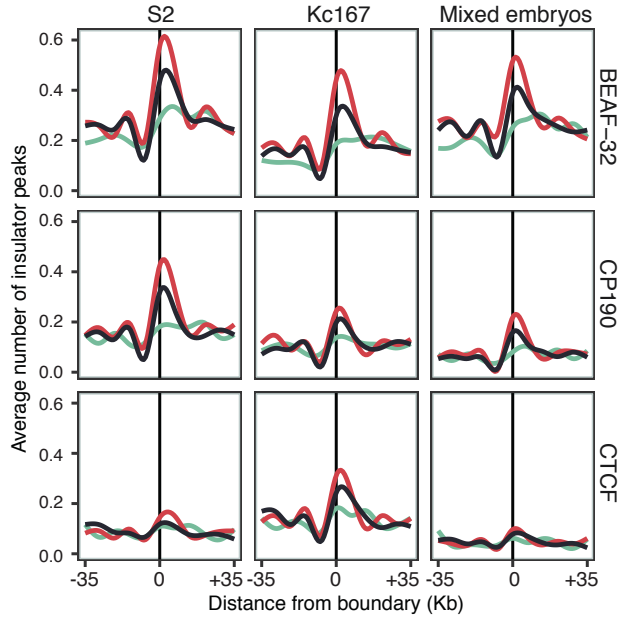

Figure 36 Adapted from Pal et al., 2018 The distribution of insulators (BEAF32, CP190, CTCF) binding peaks around the different TAD boundary classes is shown within a distance of 35Kb (10 bins distance) in S2, Kc and mixed Embryos.

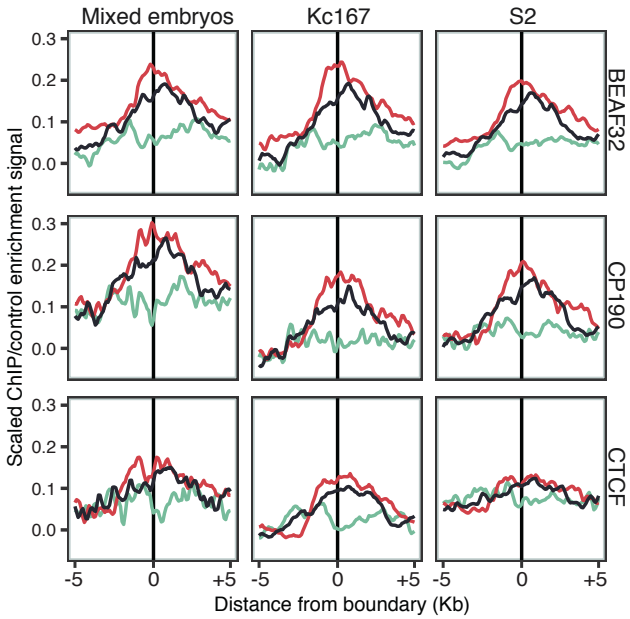

Figure 37 Adapted from Pal et al., 2018 The raw ChIP-chip enrichment signal from modENCODE is being reported. modENCODE enrichment is lowess smoothed (500bp bandwidth) M values (log2 signal intensities of CHiP over control). Enrichment values have been scaled by the 99th percentile. The average signal of insulators binding (BEAF32, CP190 and CTCF) is shown around TAD boundaries of each class.

4.18 4C-seq validates correlation between changes in accessibility and insulation

I now considered the second hypothesis. Increased accessibility of chromatin regions near dosage compensated genes lead to the changes in insulation that we observed in the male and female embryo datasets. Namely, increased accessibility will lead to an increase in the propensity of these genomic regions to participate in contacts with other nearby regions by random chance. To validate this hypothesis, I sourced nearly 200 high-quality 4C-seq datasets spread out over 29 viewpoints on the chrX, with DC induction and repression in cell lines.

I also noted that almost all of the 4C viewpoints were near to or were overlapping MSL binding sites. Of the 28 probes used, 25 (~90%) were within a distance of 3.5 Kb to a MSL binding site mid-point. I also showed in previous results that there is a distance dependency between a boundary being classified as disappearing and MSL binding sites. Considered together with previous studies (Ramírez et al. 2015) which presented the existence of a MSL binding site interaction network, lead me to hypothesize that the accessibility induced random interactions may be present in both male and female genomes, but since transcription related accessibility may be thought to be higher in the dosage compensated X chromosome, these random interactions may also increase.

Traditionally, peak calls are done using 4C-seq datasets. I argue that peak calls are designed to quantify functional interactions occurring in 4C data. 4C-seq peak callers are not designed to quantify random interactions, since these interactions would have extremely low statistical power. To quantify these random interactions, I designed 4C meta profiles for each boundary class for each probe (**Figure 38**). I observed that disappearing boundaries tend to have higher 4C signal enrichment in the regions near domain boundaries (+/- 7Kb) and also in general (**Figure 39**).

Appearing boundaries have the lowest 4C signal enrichment in general and same regions exhibit profiles which fall in-between Disappearing and Appearing boundaries (**Figure 39**). This is true for both S2 and Kc cell lines. I confirmed this pattern in all 4C probes (**Figure 40**) from the Ramírez et al., 2015 study (Ramírez et al. 2015). Although, disappearing boundaries show this same pattern in all probes across both male and female cell lines, the pattern is even greater in the male (S2) samples compared to the female (Kc) samples (**Figure 39** right, Wilcoxon p-value 0.001). My observation is consistent with a scenario where higher accessibility near disappearing borders results in lower insulation due to random contacts.

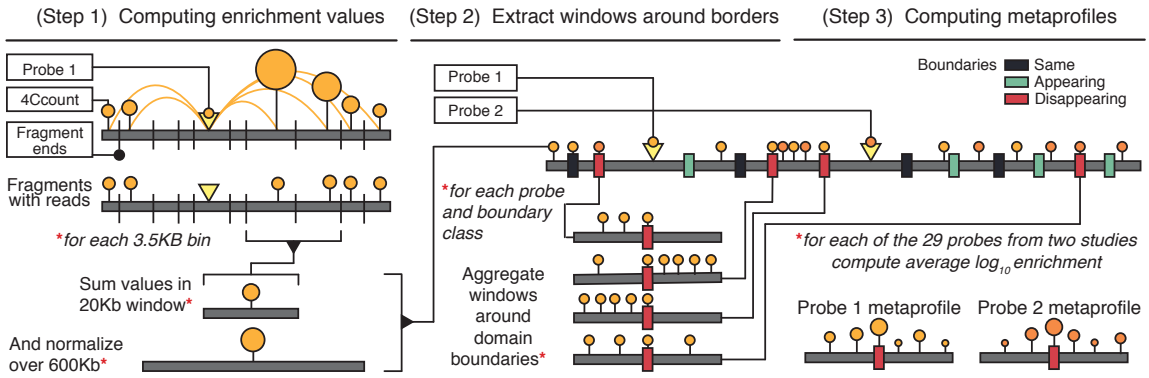

Figure 38 Adapted from Pal et al., 2018 Schematic view of the construction of a 4C meta-profile is shown. Starting from a 4C counts file, the counts are first binarised (values greater than 0 are assigned 1, everything else is 0). Taking the binarised values, these values are assigned to their corresponding 3.5Kb bin, to make the comparison possible between the 4C and Hi-C data. Then, for every 3.5Kb bin all values are summed up in a sliding window of size 20Kb. These values are then normalised by the sum of all values in a window of 600Kb. This is in line with previous literature (Ramírez et al. 2015). Taking all three boundary change classes, 35Kb windows are aggregated around all boundaries within each class and the corresponding enrichment signal is averaged and log10 transformed to create a meta-profile for a single boundary for a single 4C viewpoint.

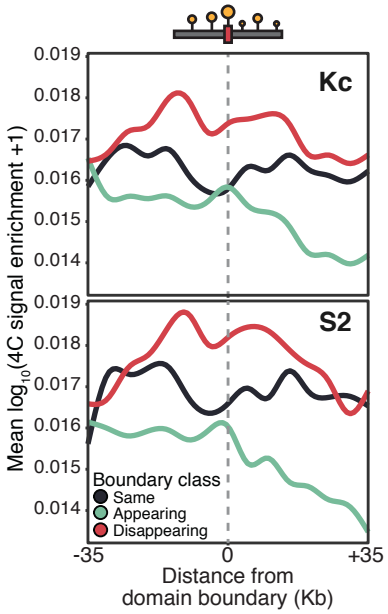

Figure 39 Adapted from Pal et al., 2018 A representative 4C meta-profile is shown for S2 and Kc cell lines. We observed, that the Disappearing boundaries show higher 4C tag enrichment than the Same and Appearing boundaries. Whereas, the Appearing boundaries show the lowest 4C tag enrichment and the Same boundaries show an intermediate signal. Notice, the average 4C tag enrichment signal peaks near (+/-7Kb) the disappearing boundaries.

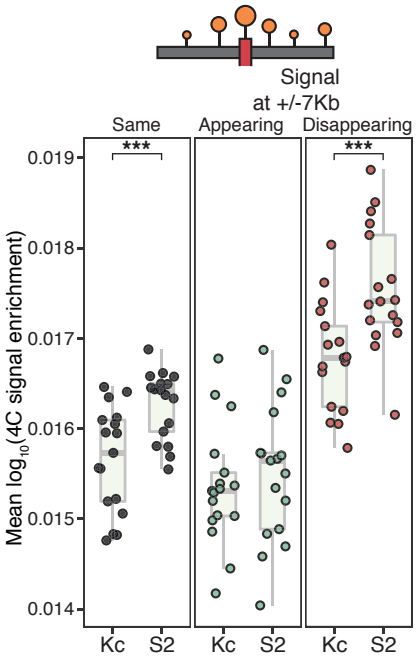

Figure 40 Adapted from Pal et al., 2018 The average signal observed from the 4C meta-profiles in the +/- 7Kb region around each TAD boundaries class is shown for all 17 probes obtained from a previous study (Ramírez et al. 2015). Each point represents the average signal observed in a +/-7Kb window around the boundaries for a single probe. We observed, that consistently the disappearing boundaries showed high signal, whereas the appearing showed the lowest signal and the same boundaries showed intermediate levels of enrichment. In case of the disappearing boundaries, the signal originating from the male S2 samples were significantly higher than the female Kc samples.

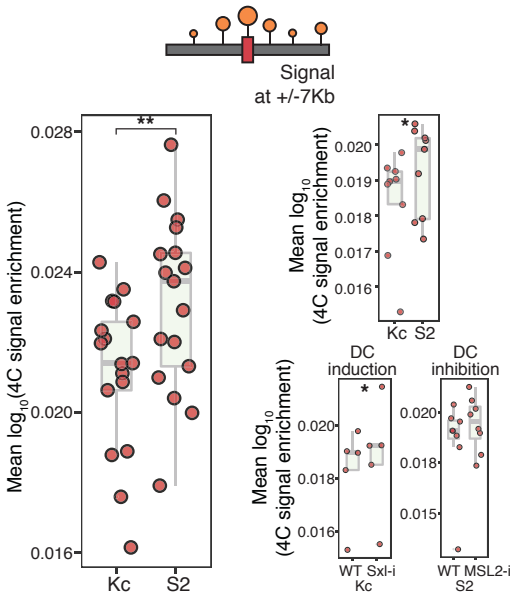

Figure 41 Adapted from Pal et al., 2018 Taking the core-set of disappearing boundaries we plot the average signal observed from the 4C meta-profiles in the +/- 7Kb region around the TAD boundaries class for 17 probes obtained from a previous study (Ramírez et al. 2015). The core-set is defined as boundaries which are lying in-between dosage compensated genes and showing at least 10% change in insulation. Each point represents the average signal observed in a +/-7Kb window around the boundaries for a single probe. Also here, we observed, that consistently the disappearing boundaries showed high signal, and the signal originating from the male S2 samples were significantly higher than the female Kc samples. Furthermore, we obtained additional 4C data for S2 and Kc samples, with induction and repression of DC (Schauer et al. 2017). We also observed comparable patterns here in the Kc vs S2 (right top, * Wilcoxon test 0.05) and the Kc vs induction of DC (right bottom).

For an even better confirmation that increased accessibility is reflected as increased 4C-seq signal, I selected a subset of disappearing boundaries located in between DC genes and showing at least 10% change in insulation. We call these boundaries the “core set” of disappearing boundaries. I then generated separate 4C meta-profiles for the core set of disappearing boundaries and found that these regions are significantly different between male and female cell lines (**Figure 41** left). This same pattern was observed across independent datasets fir Kc and S2 cell lines (**Figure 41** top right). Afterwards, I also considered 4C-seq meta-profiles profiles from cell lines with induction (silencing of Sxl) and inhibition (silencing MSL2) of DC (Schauer et al. 2017). I observed a significantly higher 4C signal around the core-set of disappearing boundaries in dosage compensated Kc (Sxl RNAi) than in control (GFP) samples (**Figure 41** bottom left). Similarly, higher 4C

signal was observed in control S2 (GFP) than in DC repressed S2 (MSL2 RNAi) cells (**Figure 41** bottom right).

4.19 Changes in CLAMP binding drive changes in insulation

I then investigated the driver of increased chromatin accessibility during dosage compensation. Previous studies have shown that CLAMP an MSL loader protein is required for proper binding of the dosage compensation complex on chrX (Soruco et al. 2013). Based on MSL dependency, CLAMP binding sites were grouped into three distinct groups. CLAMP binding at Group A sites are completely dependent on MSL binding, while group B corresponds to partial dependence and group C corresponds to independent CLAMP binding. More recently, studies have shown that CLAMP bound regions are surrounded by very large open chromatin regions which stretch upto 14Kb (Urban et al. 2017).

Therefore, I was intrigued to see if CLAMP binding is associated to disappearing domain boundaries. I found that CLAMP binding sites, specifically group A and B are enriched near disappearing boundaries (**Figure 42** right). Then I looked more closely at the core set of disappearing boundaries to validate whether a link between CLAMP and lower insulation is present. I observed that CLAMP ChIP-seq binding strength is higher near the core set of disappearing boundaries in the S2 cells as compared to Kc cells (**Figure 42** left). Therefore, it is clear that open chromatin regions near CLAMP binding sites are causing the change in insulation.

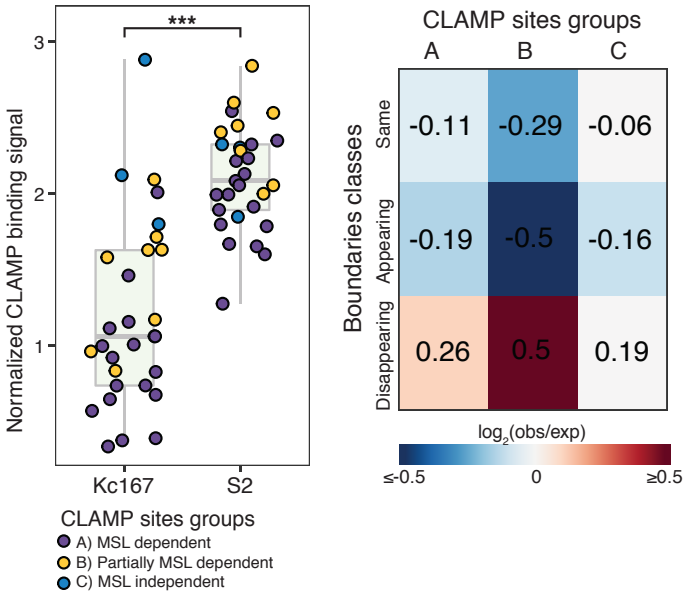

Figure 42 Adapted from Pal et al., 2018 The core-set of disappearing boundaries is strongly correlated with CLAMP. Right, Enrichment (log2 observed over expected ratio) of CLAMP binding sites around the domain borders grouped by the TAD boundary classes is shown. Three groups of CLAMP binding sites as previously defined (Soruco et al. 2013) are considered. The expected frequency was computed based on random uniform distribution of CLAMP binding sites along chrX. Left, the highest CLAMP binding strength is shown. Taking the CLAMP binding sites grouped by their affinity for dosage compensation binding, we found the point of highest wig signal intensity and call this the binding strength. The signal has been further normalised by the 99th percentile signal of the WIG file to make the comparison possible between S2 and Kc samples. The S2 shows very clearly that CLAMP has higher binding strength near the core-set of disappearing boundaries.

My observations are therefore in line with a model where local changes in chromatin accessibility associated to DC are reflected as differentially insulated regions in Hi-C maps. This study is current under review.

4.20 HiCLegos - Fast scalable solutions for analyzing Hi-C data

In our previous study, we chose well-known Hi-C datasets (Jin et al. 2013; Rao et al. 2014; Sexton et al. 2012; Dixon et al. 2012; 2015; Lieberman-Aiden et al. 2009) for benchmarking the performance of existing peak callers and TAD callers. Some of these studies, used Hi-C data binned at 5Kb (Jin et al. 2013; Rao et al. 2014). We were severely hampered in terms of processing time and memory requirements by this resolution (Forcato et al. 2017) and had to process the datasets at a much lower resolution. We resorted to reporting the results of these high-resolution datasets by binning them at 40Kb (Forcato et al. 2017). This already pointed towards a scalability issue for current Hi-C analysis tools and methods.

Furthermore, the resolution of existing Hi-C datasets have been continuously increasing over the years, with the latest highest resolution Hi-C dataset being one in mouse processed at a resolution of 850bp (Bonev et al. 2017) using the SHAMAN package (Cohen et al. 2017). Consortium efforts are also underway, with the aim of providing reference Hi-C data in cell lines alongside analysis procedures for dealing with high resolution Hi-C data (Dekker et al. 2017). Fast, efficient procedures for interactive visualisation of Hi-C data has also been proposed (Kerpedjiev et al. 2018; Durand et al. 2016).

These methods are based on python, whilst most of the biological community veers towards R. Although methods have been described for accessing Hi-C data in the R community (Lun et al. 2016; Lun and Smyth 2015), these methods are not scalable since they do not make use of on-disk data stores (Lun et al. 2016). Even when they do, these methods are not user-friendly (Lun and Smyth 2015) from a storage and access point of view. Also, there does not exist a standard data format such as GFF or Bed for the perpetuation of Hi-C data. Although efforts are underway for standardising such formats (Dekker et al. 2017). Different tools and algorithms require the usage of different data formats. The most common ones being that of an n-column tab-separated file (Durand et al. 2016) or an nom dimensional matrix file. Many pipelines adopt variations of either one leading to difficulties in porting one data format to another (Yaffe and Tanay 2011). Finally, the problem of solving dependencies on most Hi-C analysis procedures makes most of the tools and methods in-accessible to users.

I propose HiCLegos (manuscript in prep.), an R package making use of the HDF specification for storing and accessing Hi-C data. HiCLegos is a package that is integrated within the R bioconductor project (under review). Currently, HiCLegos provides methods for storing nxm dimensional matrices and mcool files generated

by the 4D nucleome project. In the access part, users are able to couple their retrieval operations with overlap operations making natural language search possible. Furthermore, HiCLegos also contains methods for making biologically relevant retrieval calls, such as the retrieval of contacts between loci separated by a certain distance (**Figure 43**). In cases, where the matrix is very large, users can decide to store only a part of the matrix until a certain distance. Although, not yet implemented, I will provide export methods for exporting Hi-C data in different formats in later releases.

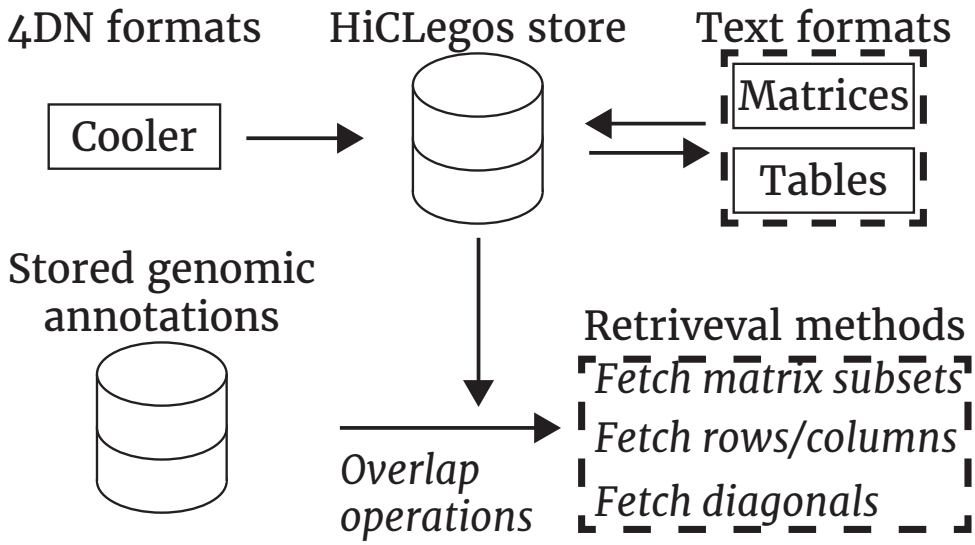

Figure 43 A schematic view of a HiCLegos workflow is shown. HiCLegos works by using on-disk HDF files. It accepts as input 4D nucleome cool files, 2D matrix files and n column tables. It further uses GenomicRanges based overlap operations to create a highly robust environment for accessing and using Hi-C data. Furthermore, it provides three basic retrieval methods, the retrieval of matrix subsets, retrieval of diagonals and the retrieval of specific rows or columns.

HiCLegos has been built so that the internal complexity of the data structure remains hidden from users, yet they are able to manipulate and access the data, while being able to keep the results of those analysis associated to the data store. As the name suggests, this package has been built so as to allow external users the ability to build additional packages using HiCLegos. As a demonstration, LSD comes packaged with HiCLegos. The time required to process data using LSD and its performance in terms of TPR and FDR have been outlined previously (see sub-

section 3.11). Finally, HiCLegos also provides single-command modules for creating Hi-C heatmaps and plotting structural features such as TADs on the maps.

To demonstrate the efficiency of this paradigm, I used HiCLegos to process high-resolution *drosophila* Hi-C data and compared it with base R procedures. HiCLegos generally outperforms base R functions in terms of read times when reading a matrix into the database (**Figure 44**). High-resolution Hi-C matrices are very demanding in terms of memory and time. I compared HiCLegos against normal R functions in terms of retrieval. HiCLegos has no additional memory overhead since the package relies on the usage of on-disk databases (**Figure 45** left). Base R generally outperforms HiCLegos methods when using low-resolution matrices, but when using high-resolution matrices HiCLegos performs faster (**Figure 45** right).

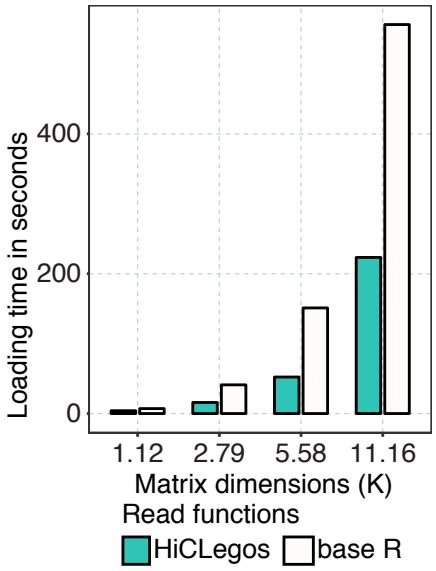

Figure 44 The data loading time of HiCLegos is compared to that of normal base R functions. I used 2D matrices to test this operation. The x-axis depicts the increasing dimensions of the matrix and the y-axis corresponds to the time required to load the matrix. As matrix size increases, HiCLegos becomes increasingly more efficient.

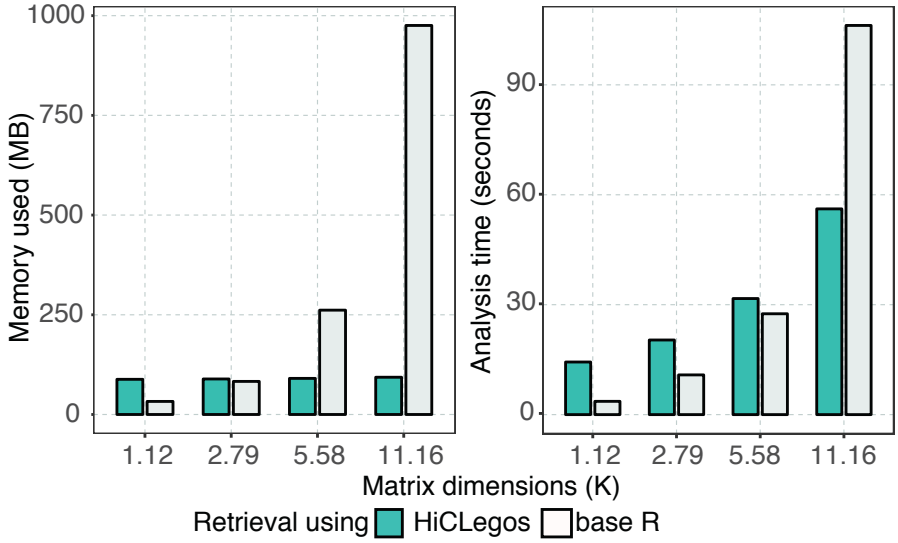

Figure 45 The efficiency of HiCLegos retrieval is depicted for retrieving matrix diagonals. HiCLegos consumes no extra memory as matrix size increases (left), whereas base R operations consume increasingly more memory. Also, as matrix size increases, the time required to complete the operation is higher for base-R than it is for HiCLegos (right).

5.0 Discussion

Higher-order chromatin structure has been studied for over a century and has been revolutionised over the past decade with the application of next-generation sequencing technologies (Dekker et al. 2002). These technical advancements have revealed a progressively compartmentalised chromatin folding landscape (Lieberman-Aiden et al. 2009; Dixon et al. 2012; Sexton et al. 2012; Phillips-Cremins et al. 2013; Rao et al. 2014; Wang et al. 2018). This landscape is highly conserved and is correlated with transcription (Rowley et al. 2017) and replication (Pope et al. 2014). It has been observed, that while compartments (Lieberman-Aiden et al. 2009) correlating with band domains (Manuelidis 1990; Saccone et al. 1993; Bernardi 1995) change during differentiation (Dixon et al. 2015), their underlying units (TADs) are highly stable and robust to change. TADs respond to experimental conditions mimicking mutational pressure by showing local changes in their structure (Rodríguez-Carballo et al. 2017; Geeven et al. 2015; Nora et al. 2017). Therefore, the reason for TAD (Topologically Associated Domain) formation and maintenance are intriguing. Previously, observations pointed towards transcription being a predictor for TAD formation (Rowley et al. 2017). Activation of transcription was later causally correlated to TAD emergence during drosophila development (Hug et al. 2017).

Dosage compensation systems provide an efficient system for studying the link between transcription and higher-order chromatin structure. In mammalian dosage compensation systems where one X chromosome is completely silenced, TADs are not observed (Giorgetti et al. 2016). Indeed, the mammalian inactivated X chromosome expresses a few genes, and the regions where these genes are expressed, TAD like structures are observed (Giorgetti et al. 2016). Together these results suggest a causal link between transcription and chromatin structure. On the

other hand, TADs are known to be bounded by insulator binding sites (Sexton et al. 2012; Hou et al. 2012; Dixon et al. 2012). These insulator binding sites are directionally oriented in an inward facing manner (Rao et al. 2014; Guo et al. 2015). Mutations and changes at these sites have been implicated in disease biology. Notably, the IDH mutations are considered to be an oncogenic driver in cancer (Cohen et al. 2013). In IDH mutant cells, methylation in CTCF sites near the PDGFRA TAD lead to a lower insulation of the TAD and an increase in contacts between PDGFRA, a cancer driver, and enhancers outside the TAD (Flavahan et al. 2016). A change in orientation of CTCF sites has been linked to limb malformation (Lupiáñez et al. 2015) and changes in gene expression patterns (Guo et al. 2015). Therefore, higher-order chromatin structure cannot be ruled out as a by-product of processes such as transcription and replication. Broadly viewed, these results point towards a causal and maintainer relationship between transcription and higher-order chromatin folding.

Herein, the dosage compensation mechanisms in *D. melanogaster* and *C. elegans* comes into focus. Both mechanisms affect fine grain changes in transcriptional regulation. In *drosophila*, active genes on the single-copy male chrX are up-regulated by an average factor of two-fold. On the other hand, dosage compensation in *C. elegans* affects a down-regulation of both X chromosomes in the hermaphrodite. In case of the latter, the X chromosomes adopt a distinct structure which is more insulated than the autosomes (Crane et al. 2015). Yet, in case of the former, i.e. *drosophila*, genome-wide studies using Hi-C did not observe any structural changes (Ramírez et al. 2015; Schauer et al. 2017). Firstly, previous studies using the same Hi-C datasets reported no structural differences (Schauer et al. 2017) due to the partitioning of the genome into compartments (Lieberman-Aiden et al. 2009). Compartment analysis (Lieberman-Aiden et al. 2009) has revealed

structural changes during differentiation (Dixon et al. 2015), where various transcriptional networks are silenced and other activated. The same cannot be expected in *drosophila* dosage compensation where already active genes are up-regulated. Secondly, prior studies (Ramírez et al. 2015) also utilised drosophila cell lines for the analysis. *Drosophila* cell lines, principally the S2 (male) and Kc (female) cell lines are severely biased by copy number changes (Lee et al. 2014). These copy number differences tend to influence the Hi-C signal, which is not accounted for by current implicit normalisation procedures (Servant et al. 2018). On top of this, wild-type *drosophila* males also carry one copy of chrX, presenting half as many reads as its counterpart in females. This requires the adoption of chromosome specific analysis and normalisation procedure. Lastly, a scenario where selected genes are affected by dosage compensation would not affect the chromatin fibre in the same way across the entire chromosome. Most TAD calling procedures are not designed for detecting local fluctuations in the chromatin fibre, rather these algorithms detect large-scale folding structures (Dixon et al. 2012; Serra et al. 2016). Therefore, a need exists to investigate local genome compartmentalisation further.

I developed ad-hoc analysis procedures allowing me to detect a change in the global interaction landscape in male flies. The non-parametric selection of top-scoring interactions and polymer folding simulations helped me to confirm my observations. The differences in interaction decay between the dosage compensated X chromosome and autosomes were small, but robust, reproducible and significant. I concluded that the male chrX is more prone to participate in long-range contacts. These long-range contacts did not cluster together, suggesting that these interactions were not stable interactions as would be expected from functionally relevant interactions but rather random events occurring due to increased accessibility. My observations pointed towards a globally more open and

accessible chrX, resulting in more Hi-C signal at larger distances.

I then moved on to investigate the differences in chrX domains between sexes. For this task, I developed algorithms to take into account the differential dosage compensation effects in the genome and to also take into account copy number related issues. I created a TAD calling procedure, Local Score Differentiator (LSD) which uses locally defined thresholds. This, as opposed to using genome-wide or chromosome-wise defined thresholds ensures that LSD is sensitive to local fluctuations in the chromatin fibre. LSD is also a very fast boundary calling procedure. This allowed us to utilise high-resolution matrices and to identify domain boundaries that change between male and female. I then identified a subset of lowly insulated domain boundaries which are associated to dosage compensation complex binding and transcriptional response of genes to dosage compensation. To confirm lower levels of insulation in chrX, I utilised publicly available 4C-seq data across cell lines, including data for induction (in female cells) or inhibition (in male cells) of DC. Thus, I concluded that changes in chromatin accessibility is also affecting insulation across these boundaries and that these changes are detected by Hi-C.

I then investigated the general reason preceding this change in accessibility. I looked into insulator binding profiles across cell lines. Although specific patterns were visible, a clear generalised pattern was not observed. Analysing known dosage compensation co-factors allowed me to identify differences in CLAMP binding. CLAMP is a protein implicated in the binding of the dosage compensation complex (Soruco et al. 2013). Recently, it was reported that CLAMP binding leads to an increase in chromatin accessibility (Urban et al. 2017). This clearly explained the preservation of insulator binding alongside localised changes in insulation seen in the Hi-C matrices.

In this project I have shown that the dosage compensation of fly chrX leads to an increase in global accessibility and local changes in insulation. By developing new analysis methods, I have been able to detect these changes with genome-wide Hi-C data. This is in line with previous literature, which postulated that an increase in accessibility may be expected. I have shown that these structural changes in 3D chromatin architecture are subtle but detectable with Hi-C. This the first such report of its kind as previous literature utilising Hi-C data on *drosophila* cell lines were not able to detect these changes.

With the help of this particular project and our previous work where we comparatively assessed the performance of Hi-C analysis procedures in terms of peak calling and TAD calling, I identified a need for a standard Hi-C analysis framework within the R community which is predominantly the language of choice for biologists. The variety of Hi-C data formats makes analysis of Hi-C data seemingly complicated and time intensive. Consortium efforts are currently underway for standardising Hi-C data formats (Dekker et al. 2017). Yet, such formats may not play well with pre-existing Hi-C analysis pipelines and methods. This may seem to be a triviality as it is only a matter of re-casting the data into the format of choice. But as Hi-C data generation achieves newer heights, the time required to re-cast this data also increases. The highest resolution Hi-C data generated nearly 40 billion reads and analysed the dataset at a resolution of 850bp in a mouse genome (Bonev et al. 2017). Therefore, it is not exaggerated to state that the time is near when Hi-C data binned at 500bp in humans is the norm. Indeed, during the course of our previous study (Forcato et al. 2017), we encountered unoptimised code which created severe bottlenecks in hicpipe (Yaffe and Tanay 2011) when analysing high-resolution Hi-C data. By changing a single line, we were able to achieve significantly faster processing times. These high-resolution Hi-C

datasets are extremely difficult to access, as these datasets require a long time to load. For analysing such datasets the usage of on-disk data formats has been proposed. HDF, or Hierarchical Data Format is one such on-disk data format. This is a general specification and therefore it requires adaption towards specific use cases. HDF data formats are in use in the Python ecosystem but Hi-C analysis libraries based on HDF files are lacking in the R statistical environment. To meet this requirement I have developed HiCLegos. HiCLegos, as the name suggests is a library which aims to be a building block for future Hi-C analysis tools and methods utilising on-disk data formats in the R ecosystem. The library encapsulates the underlying complexity of the HDF specification and exposes biologically meaningful data access methods. An example of such a method is the retrieval of values corresponding to interactions between genomic loci separated by a certain distance. By providing users the ability to retrieve Hi-C data using human readable genomic coordinates, HiCLegos makes the entry into Hi-C data analysis easier for beginners. Furthermore, Hi-C legos simplifies data loading by providing specific methods for loading matrices, tables or binary data formats created with other such libraries in the python ecosystem. As a proof of concept, Local Score Differentiator (LSD) also uses HiCLegos methods for data access making it an extremely fast TAD calling procedure. HiCLegos also contains visualisation methods for plotting publication-ready heat maps. In total, HiCLegos represents nearly 3400 lines of code, spread out across 177 unique functions, of which 35 are exposed to the user.

In summary, novel analysis methods and frameworks allowed us to investigate the changing structure of the *drosophila* dosage compensated X chromosome. Using these tools, we found that as a whole increased chromatin accessibility affects Hi-C signal and local chromatin compartmentalisation. Although, the reasons behind the specific structural changes driving this change in

insulation is still unclear. Essentially the question remains what structural events lead up to an decrease in insulation, and how do these events connect the preferential positioning of CLAMP binding sites and gene TSS near lowly insulated regions. Emerging hypothesis within the field suggests that transcription induced supercoiling may play a role in driving genome compartmentalisation (Racko et al. 2017). Furthermore, the structural events preceding or following gene up-regulation is still not clear. Previous studies have postulated that increased recycling of polymerase via gene looping and/or increased processivity (i.e. decreased premature termination) may be possible mechanisms for gene up-regulation (Ferrari et al. 2014).

Previous studies have shown that dosage compensation binding sites tend to contact each other and possibly occupy a spatially distinct region in 3D space (Ramírez et al. 2015). Yet, these regions interact in a similar manner in both the male and female genomes (Ramírez et al. 2015), therefore their contribution towards a changing chromosome conformation was unlikely. Although we cannot rule out their contribution in changing global chromatin conformation, our preliminary results suggest that lowly insulated regions tend to occupy a very distinct region in the nuclear space. This shows a very striking difference between the male and female genomes. Taking into context the findings from our study, I hypothesise that these regions form transcriptional hubs similar to the active chromatin hub observed in the β-globin locus (Gavrilov et al. 2013). Since, these low-insulation regions are also near MSL binding sites and CLAMP binding sites, this hypothesis requires further investigation.

6.0 References

Alekseyenko AA, Peng S, Larschan E, Gorchakov AA, Lee O-K, Kharchenko P, McGrath SD, Wang CI, Mardis ER, Park PJ, et al. 2008. A sequence motif within chromatin entry sites directs MSL establishment on the Drosophila X chromosome. *Cell* **134**: 599–609.

Ay F, Bailey TL, Noble WS. 2014. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Research* **24**: 999–1011.

Baù D, Sanyal A, Lajoie BR, Capriotti E, Byron M, Lawrence JB, Dekker J, Marti-Renom MA. 2011. The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. *Nature Publishing Group* **18**: 107–114.

Beagrie RA, Scialdone A, Schueler M, Kraemer DCA, Chotalia M, Xie SQ, Barbieri M, de Santiago I, Lavitas L-M, Branco MR, et al. 2017. Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* **543**: 519–524.

Belton J-M, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. 2012. Hi–C: A comprehensive technique to capture the conformation of genomes. *Methods* **58**: 268–276.

Berezney R, Coffey DS. 1977. Nuclear matrix. Isolation and characterization of a framework structure from rat liver nuclei. *J Cell Biol* **73**: 616–637.

Berlivet S, Paquette D, Dumouchel A, Langlais D, Dostie J, Kmita M. 2013. Clustering of tissue-specific sub-TADs accompanies the regulation of HoxA genes in developing limbs. ed. B. Ren. *PLoS Genet* **9**: e1004018.

Bernardi G. 1995. The Human Genome: Organization and Evolutionary History. *Annu Rev Genet* **29**: 445–476.

Blobel G. 1985. Gene gating: a hypothesis. *Proceedings of the National Academy of Sciences of the United States of America* **82**: 8527–8529.

Bonev B, Mendelson Cohen N, Szabo Q, Fritsch L, Papadopoulos GL, Lubling Y, Xu X, Lv X, Hugnot J-P, Tanay A, et al. 2017. Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell* **171**: 557–572.e24.

Carrivain P, Barbi M, Victor J-M. 2014. In silico single-molecule manipulation of DNA with rigid body dynamics. ed. S.-J. Chen. **10**: e1003456.

Charlesworth B. 2001. Genome analysis: More Drosophila Y chromosome genes. *Curr Biol* **11**: R182–4.

Cohen AL, Holmen SL, Colman H. 2013. IDH1 and IDH2 mutations in gliomas. *Curr Neurol Neurosci Rep* **13**: 345.

Cohen NM, Olivares-Chauvet P, Lubling Y, Baran Y, Lifshitz A, Hoichman M, Tanay A. 2017. SHAMAN: bin-free randomization, normalization and screening of

Hi-C matrices. *bioRxiv* 187203.

Comings DE. 1968. The rationale for an ordered arrangement of chromatin in the interphase nucleus. *Am J Hum Genet* **20**: 440–460.

Contrino S, Smith RN, Butano D, Carr A, Hu F, Lyne R, Rutherford K, Kalderimis A, Sullivan J, Carbon S, et al. 2012. modMine: flexible access to modENCODE data. *Nucleic Acids Res* **40**: D1082–8.

Cournac A, Marie-Nelly H, Marbouty M, Koszul R, Mozziconacci J. 2012. Normalization of a chromosomal contact map. *BMC genomics* **13**: 436.

Crane E, Bian Q, McCord RP, Lajoie BR, Wheeler BS, Ralston EJ, Uzawa S, Dekker J, Meyer BJ. 2015. Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature* **523**: 240–244.

Cremer T, Cremer C. 2006. Rise, fall and resurrection of chromosome territories: a historical perspective. Part I. The rise of chromosome territories. *Eur J Histochem* **50**: 161–176.

Cremer T, Cremer C, Baumann H, Luedtke EK, Sperling K, Teuber V, Zorn C. 1982. Rabl's model of the interphase chromosome arrangement tested in Chinise hamster cells by premature chromosome condensation and laser-UV-microbeam experiments. *Human Genetics* **60**: 46–56.

Cremer T, Cremer M. 2010. Chromosome territories. *Cold Spring Harb Perspect Biol* **2**: a003889–a003889.

Cremer T, Landegent J, Brückner A, Scholl HP, Schardin M, Hager HD, Devilee P, Pearson P, van der Ploeg M. 1986. Detection of chromosome aberrations in the human interphase nucleus by visualization of specific target DNAs with radioactive and non-radioactive in situ hybridization techniques: diagnosis of trisomy 18 with probe L1.84. *Human Genetics* **74**: 346–352.

Cuny G, Soriano P, (null) GM, Bernardi G. 1981. The major components of the mouse and human genomes: 1. Preparation, basic properties and compositional heterogeneity. *Eur J Biochem* **115**: 227–233.

Darrow EM, Huntley MH, Dudchenko O, Stamenova EK, Durand NC, Sun Z, Huang S-C, Sanborn AL, Machol I, Shamim M, et al. 2016. Deletion of DXZ4 on the human inactive X chromosome alters higher-order genome architecture. *Proceedings of the National Academy of Sciences of the United States of America* **113**: E4504–12.

de Wit E, de Laat W. 2012. A decade of 3C technologies: insights into nuclear organization. *Genes & Development* **26**: 11–24.

Dekker J, Belmont AS, Guttman M, Leshyk VO, Lis JT, Lomvardas S, Mirny LA, O'Shea CC, Park PJ, Ren B, et al. 2017. The 4D nucleome project. *Nature* **549**: 219–226.

Dekker J, Heard E. 2015. Structural and functional diversity of Topologically Associating Domains. *FEBS letters* **589**: 2877–2884.

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

Dhar V, Skoultchi AI, Schildkraut CL. 1989. Activation and repression of a beta-globin gene in cell hybrids is accompanied by a shift in its temporal replication. *Mol Cell Biol* **9**: 3524–3532.

Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, Ye Z, Kim A, Rajagopal N, Xie W, et al. 2015. Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**: 331–336.

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

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.

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

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

Duan Z, Andronescu M, Schutz K, Lee C, Shendure J, Fields S, Noble WS, Anthony Blau C. 2012. A genome-wide 3C-method for characterizing the three-dimensional architectures of genomes. *Methods* **58**: 277–288.

Duan Z, Andronescu M, Schutz K, McIlwain S, Kim YJ, Lee C, Shendure J, Fields S, Blau CA, Noble WS. 2010. A three-dimensional model of the yeast genome. *Nature* **465**: 363–367.

Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, Aiden EL. 2016. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**: 95–98.

Fabre PJ, Leleu M, Mormann BH, Lopez-Delisle L, Noordermeer D, Beccari L, Duboule D. 2017. Large scale genomic reorganization of topological domains at the HoxD locus. *Genome Biology* **18**: 149.

Feldherr CM, Kallenbach E, Schultz N. 1984. Movement of a karyophilic protein through the nuclear pores of oocytes. *J Cell Biol* **99**: 2216–2222.

Ferrari F, Alekseyenko AA, Park PJ, Kuroda MI. 2014. Transcriptional control of a whole chromosome: emerging models for dosage compensation. *Nature Structural Molecular Biology* **21**: 118–125.

Ferrari F, Plachetka A, Alekseyenko AA, Jung YL, Ozsolak F, Kharchenko PV,

Park PJ, Kuroda MI. 2013. “Jump start and gain” model for dosage compensation in Drosophila based on direct sequencing of nascent transcripts. *Cell Reports* **5**: 629–636.

Filippova D, Patro R, Duggal G, Kingsford C. 2014. Identification of alternative topological domains in chromatin. *Algorithms Mol Biol* **9**: 14.

Flavahan WA, Drier Y, Liau BB, Gillespie SM, Venteicher AS, Stemmer-Rachamimov AO, Suvà ML, Bernstein BE. 2016. Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* **529**: 110–114.

Forcato M, Nicoletti C, Pal K, Livi CM, Ferrari F, Bicciato S. 2017. Comparison of computational methods for Hi-C data analysis. *Nat Meth* **14**: 679–685.

Fudenberg G, Imakaev M, Lu C, Goloborodko A, Abdennur N, Mirny LA. 2016. Formation of Chromosomal Domains by Loop Extrusion. *Cell Reports* **15**: 2038–2049.

Gavrilov AA, Gushchanskaya ES, Strelkova O, Zhironkina O, Kireev II, Iarovaia OV, Razin SV. 2013. Disclosure of a structural milieu for the proximity ligation reveals the elusive nature of an active chromatin hub. *Nucleic Acids Res* **41**: 3563–3575.

Geeven G, Zhu Y, Kim BJ, Bartholdy BA, Yang S-M, Macfarlan TS, Gifford WD, Pfaff SL, Verstegen MJAM, Pinto H, et al. 2015. Local compartment changes and regulatory landscape alterations in histone H1-depleted cells. *Genome Biology* **16**: 289.

Gerace L, Blum A, Blobel G. 1978. Immunocytochemical localization of the major polypeptides of the nuclear pore complex-lamina fraction. Interphase and mitotic distribution. *J Cell Biol* **79**: 546–566.

Giorgetti L, Galupa R, Nora EP, Piolot T, Lam F, Dekker J, Tiana G, Heard E. 2014. Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell* **157**: 950–963.

Giorgetti L, Lajoie BR, Carter AC, Attia M, Zhan Y, Xu J, Chen CJ, Kaplan N, Chang HY, Heard E, et al. 2016. Structural organization of the inactive X chromosome in the mouse. *Nature* **535**: 575–579.

Grimaud C, Becker PB. 2009. The dosage compensation complex shapes the conformation of the X chromosome in Drosophila. *Genes & Development* **23**: 2490–2495.

Guo Y, Xu Q, Canzio D, Shou J, Li J, Gorkin DU, Jung I, Wu H, Zhai Y, Tang Y, et al. 2015. CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell* **162**: 900–910.

Haddad N, Vaillant C, Jost D. 2017. IC-Finder: inferring robustly the hierarchical organization of chromatin folding. *Nucleic Acids Res* **45**: e81.

Hnisz D, Weintraub AS, Day DS, Valton A-L, Bak RO, Li CH, Goldmann J, Lajoie

BR, Fan ZP, Sigova AA, et al. 2016. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science (New York, NY)* **351**: 1454–1458.

Holmquist GP. 1989. Evolution of chromosome bands: molecular ecology of noncoding DNA. *J Mol Evol* **28**: 469–486.

Hou C, Li L, Qin ZS, Corces VG. 2012. Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. *Molecular Cell* **48**: 471–484.

Hsieh T-HS, Fudenberg G, Goloborodko A, Rando OJ. 2016. Micro-C XL: assaying chromosome conformation from the nucleosome to the entire genome. *Nat Meth* **13**: 1009–1011.

Hsieh T-HS, Weiner A, Lajoie B, Dekker J, Friedman N, Rando OJ. 2015. Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell* **162**: 108–119.

Hu M, Deng K, Selvaraj S, Qin Z, Ren B, Liu JS. 2012. HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics* **28**: 3131–3133.

Hug CB, Grimaldi AG, Kruse K, Vaquerizas JM. 2017. Chromatin Architecture Emerges during Zygotic Genome Activation Independent of Transcription. *Cell* **169**: 216–228.e19.

Hutchison N, Weintraub H. 1985. Localization of DNAase I-sensitive sequences to specific regions of interphase nuclei. *Cell* **43**: 471–482.

Hwang Y-C, Lin C-F, Valladares O, Malamon J, Kuksa PP, Zheng Q, Gregory BD, Wang L-S. 2015. HIPPIE: a high-throughput identification pipeline for promoter interacting enhancer elements. *Bioinformatics* **31**: 1290–1292.

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

Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, Yen C-A, Schmitt AD, Espinoza CA, Ren B. 2013. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**: 290–294.

Joyce EF, Erceg J, Wu C-T. 2016. Pairing and anti-pairing: a balancing act in the diploid genome. *Curr Opin Genet Dev* **37**: 119–128.

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

Kerpedjiev P, Abdennur N, Lekschas F, McCallum C, Dinkla K, Strobelt H, Luber JM, Ouellette SB, Azhir A, Kumar N, et al. 2018. HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome Biology* **19**: 125.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat*

Meth **9**: 357–359.

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

Lawrence JB, Villnave CA, Singer RH. 1988. Sensitive, high-resolution chromatin and chromosome mapping in situ: presence and orientation of two closely integrated copies of EBV in a lymphoma line. *Cell* **52**: 51–61.

Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. 2013. Software for computing and annotating genomic ranges. ed. A. Prlic. **9**: e1003118.

Le Dily F, Baù D, Pohl A, Vicent GP, Serra F, Soronellas D, Castellano G, Wright RHG, Ballare C, Filion G, et al. 2014. Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. *Genes & Development* **28**: 2151–2162.

Lee H, McManus CJ, Cho D-Y, Eaton M, Renda F, Somma MP, Cherbas L, May G, Powell S, Zhang D, et al. 2014. DNA copy number evolution in Drosophila cell lines. *Genome Biology* **15**: R70.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Li L, Lyu X, Hou C, Takenaka N, Nguyen HQ, Ong C-T, Cubeñas-Potts C, Hu M, Lei EP, Bosco G, et al. 2015. Widespread rearrangement of 3D chromatin organization underlies polycomb-mediated stress-induced silencing. *Molecular Cell* **58**: 216–231.

Lichter P, Cremer T, Borden J, Manuelidis L, Ward DC. 1988a. Delineation of individual human chromosomes in metaphase and interphase cells by in situ suppression hybridization using recombinant DNA libraries. *Human Genetics* **80**: 224–234.

Lichter P, Cremer T, Tang CJ, Watkins PC, Manuelidis L, Ward DC. 1988b. Rapid detection of human chromosome 21 aberrations by in situ hybridization. *Proceedings of the National Academy of Sciences of the United States of America* **85**: 9664–9668.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**: 289–293.

Lois R, Freeman L, Villeponteau B, Martinson HG. 1990. Active beta-globin gene transcription occurs in methylated, DNase I-resistant chromatin of nonerythroid chicken cells. *Mol Cell Biol* **10**: 16–27.

Lun ATL, Perry M, Ing-Simmons E. 2016. Infrastructure for genomic interactions: Bioconductor classes for Hi-C, ChIA-PET and related experiments. *F1000Res* **5**: 950.

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

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

Manuelidis L. 1990. A view of interphase chromosomes. *Science* **250**: 1533–1540.

Manuelidis L. 1985. Individual interphase chromosome domains revealed by in situ hybridization. *Human Genetics* **71**: 288–293.

Manuelidis L, Borden J. 1988. Reproducible compartmentalization of individual chromosome domains in human CNS cells revealed by in situ hybridization and three-dimensional reconstruction. *Chromosoma* **96**: 397–410.

Mifsud B, Martincorena I, Darbo E, Sugar R, Schoenfelder S, Fraser P, Luscombe NM. 2017. GOTHiC, a probabilistic model to resolve complex biases and to identify real interactions in Hi-C data. *PloS One* **12**: e0174744.

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

Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, Laue ED, Tanay A, Fraser P. 2013. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**: 59–64.

Nagano T, Lubling Y, Yaffe E, Wingett SW, Dean W, Tanay A, Fraser P. 2015. Single-cell Hi-C for genome-wide detection of chromatin interactions that occur simultaneously in a single cell. *Nature Protocols* **10**: 1986–2003.

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

Noordermeer D, Branco MR, Splinter E, Klous P, van Ijcken W, Swagemakers S, Koutsourakis M, van der Spek P, Pombo A, de Laat W. 2008. Transcription and chromatin organization of a housekeeping gene cluster containing an integrated beta-globin locus control region. ed. J.T. Lee. *PLoS Genet* **4**: e1000016.

Noordermeer D, de Wit E, Klous P, van de Werken H, Simonis M, Lopez-Jones M, Eussen B, de Klein A, Singer RH, de Laat W. 2011a. Variegated gene expression caused by cell-specific long-range DNA interactions. *Nat Cell Biol* **13**: 944–951.

Noordermeer D, Leleu M, Splinter E, Rougemont J, de Laat W, Duboule D. 2011b. The dynamic architecture of Hox gene clusters. *Science (New York, NY)* **334**: 222–225.

Nora EP, Goloborodko A, Valton A-L, Gibcus JH, Uebersohn A, Abdennur N, Dekker J, Mirny LA, Bruneau BG. 2017. Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell* **169**: 930–944.e22.

Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, van Berkum NL, Meisig J, Sedat J, et al. 2012. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**: 381–385.

Ogiyama Y, Schuettengruber B, Papadopoulos GL, Chang J-M, Cavalli G. 2018. Polycomb-Dependent Chromatin Looping Contributes to Gene Silencing during Drosophila Development. *Molecular Cell* **71**: 73–.

Palstra RJ, Tolhuis B, Splinter E, Nijmeijer R, Grosveld F, de Laat W. 2003. The beta-globin nuclear compartment in development and erythroid differentiation. *Nature Genetics* **35**: 190–194.

Phillips-Cremins JE, Sauria MEG, Sanyal A, Gerasimova TI, Lajoie BR, Bell JSK, Ong C-T, Hookway TA, Guo C, Sun Y, et al. 2013. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* **153**: 1281–1295.

Pinkel D, Landegent J, Collins C, Fuscoe J, Segraves R, Lucas J, Gray J. 1988. Fluorescence in situ hybridization with human chromosome-specific libraries: detection of trisomy 21 and translocations of chromosome 4. *Proceedings of the National Academy of Sciences of the United States of America* **85**: 9138–9142.

Pinkel D, Straume T, Gray JW. 1986. Cytogenetic analysis using quantitative, high-sensitivity, fluorescence hybridization. *Proceedings of the National Academy of Sciences of the United States of America* **83**: 2934–2938.

Pope BD, Ryba T, Dileep V, Yue F, Wu W, Denas O, Vera DL, Wang Y, Hansen RS, Canfield TK, et al. 2014. Topologically associating domains are stable units of replication-timing regulation. *Nature* **515**: 402–405.

Ptashne M. 1986. Gene regulation by proteins acting nearby and at a distance. *Nature* **322**: 697–701.

Racko D, Benedetti F, Dorier J, Stasiak A. 2017. Transcription-induced supercoiling as the driving force of chromatin loop extrusion during formation of TADs in interphase chromosomes. *Nucleic Acids Res*.

Ramírez F, Bhardwaj V, Arrigoni L, Lam KC, Grüning BA, Villaveces J, Habermann B, Akhtar A, Manke T. 2018. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nature Communications* **9**: 189.

Ramírez F, Lingg T, Toscano S, Lam KC, Georgiev P, Chung H-R, Lajoie BR, 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. *Molecular Cell* **60**: 146–162.

Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**: W160–5.

Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, et al. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**: 1665–1680.

Rodríguez-Carballo E, Lopez-Delisle L, Zhan Y, Fabre PJ, Beccari L, El-Idrissi I, Huynh THN, Ozadam H, Dekker J, Duboule D. 2017. The HoxD cluster is a dynamic and resilient TAD boundary controlling the segregation of antagonistic regulatory landscapes. *Genes & Development* **31**: 2264–2281.

Rowley MJ, Nichols MH, Lyu X, Ando-Kuri M, Rivera ISM, Hermetz K, Wang P, Ruan Y, Corces VG. 2017. Evolutionarily Conserved Principles Predict 3D Chromatin Organization. *Molecular Cell* **67**: 837–852.e7.

Rudan MV, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A, Hadjur S. 2015. Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. *Cell Reports* **10**: 1297–1309.

Saccone S, Cacciò S, Kusuda J, Andreozzi L, Bernardi G. 1996. Identification of the gene-richest bands in human chromosomes. *Gene* **174**: 85–94.

Saccone S, De Sario A, Wiegant J, Raap AK, Valle Della G, Bernardi G. 1993. Correlations between isochores and chromosomal bands in the human genome. *Proceedings of the National Academy of Sciences of the United States of America* **90**: 11929–11933.

Sadoni N, Langer S, Fauth C, Bernardi G, Cremer T, Turner BM, Zink D. 1999. Nuclear organization of mammalian genomes. Polar chromosome territories build up functionally distinct higher order compartments. *J Cell Biol* **146**: 1211–1226.

Samata M, Akhtar A. 2018. Dosage Compensation of the X Chromosome: A Complex Epigenetic Assignment Involving Chromatin Regulators and Long Noncoding RNAs. *Annu Rev Biochem* **87**: annurev–biochem–062917–011816–350.

Schauer T, Ghavi-Helm Y, Sexton T, Albig C, Regnard C, Cavalli G, Furlong EE, Becker PB. 2017. Chromosome topology guides the Drosophila Dosage Compensation Complex for target gene activation. *EMBO Rep* **18**: 1854–1868.

Serra F, Baù D, Filion G, Marti-Renom MA. 2016. *Structural features of the fly chromatin colors revealed by automatic three-dimensional modeling.* Cold Spring Harbor Labs Journals.

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

Servant N, Varoquaux N, Heard E, Barillot E, Vert J-P. 2018. Effective normalization for copy number variation in Hi-C data. *BMC Bioinformatics* **19**: 313.

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

Shih H-Y, Krangel MS. 2013. Chromatin architecture, CCCTC-binding factor, and V(D)J recombination: managing long-distance relationships at antigen receptor loci. *J Immunol* **190**: 4915–4921.

Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, van Steensel B, de Laat W. 2006. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature Genetics* **38**: 1348–1354.

Simonis M, Kooren J, de Laat W. 2007. An evaluation of 3C-based methods to capture DNA interactions. *Nat Meth* **4**: 895–901.

Soruco MML, Chery J, Bishop EP, Siggers T, Tolstorukov MY, Leydon AR, Sugden AU, Goebel K, Feng J, Xia P, et al. 2013. The CLAMP protein links the MSL complex to the X chromosome during Drosophila dosage compensation. *Genes & Development* **27**: 1551–1556.

Spill YG, Castillo D, Marti-Renom MA. 2017. Binless normalization of Hi-C data provides significant interaction and difference detection independently of resolution. *bioRxiv* 214403.

Splinter E, de Wit E, Nora EP, Klous P, van de Werken HJG, Zhu Y, Kaaij LJT, van IJcken W, Gribnau J, Heard E, et al. 2011. The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. *Genes & Development* **25**: 1371–1383.

Stack SM, Brown DB, Dewey WC. 1977. Visualization of interphase chromosomes. *J Cell Sci* **26**: 281–299.

Straub T, Grimaud C, Gilfillan GD, Mitterweger A, Becker PB. 2008. The chromosomal high-affinity binding sites for the Drosophila dosage compensation complex. *PLoS Genet* **4**: e1000302.

Sun FL, Cuaycong MH, Craig CA, Wallrath LL, Locke J, Elgin SC. 2000. The fourth chromosome of Drosophila melanogaster: interspersed euchromatic and heterochromatic domains. *Proceedings of the National Academy of Sciences of the United States of America* **97**: 5340–5345.

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al. 2012. The accessible chromatin landscape of the human genome. *Nature* **489**: 75–82.

Tolhuis B, Palstra RJ, Splinter E, Grosveld F, de Laat W. 2002. Looping and interaction between hypersensitive sites in the active beta-globin locus. *Molecular Cell* **10**: 1453–1465.

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

Villa R, Schauer T, Smialowski P, Straub T, Becker PB. 2016. PionX sites mark the X chromosome for dosage compensation. *Nature* **537**: 244–248.

Vogel F, Schroeder TM. 1974. The internal order of the interphase nucleus. *Humangenetik* **25**: 265–297.

Wang Q, Sun Q, Czajkowsky DM, Shao Z. 2018. Sub-kb Hi-C in D. melanogaster reveals conserved characteristics of TADs between insect and mammalian cells. *Nature Communications* **9**: 331.

Weinreb C, Raphael BJ. 2015. Identification of hierarchical chromatin domains. *Bioinformatics* btv485.

Wischnitzer S. 1973. The submicroscopic morphology of the interphase nucleus. *Int Rev Cytol* **34**: 1–48.

Yaffe E, Tanay A. 2011. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature Genetics* **43**: 1059–1065.

Yang T, Zhang F, Yardımcı GG, Song F, Hardison RC, Noble WS, Yue F, Li Q. 2017. HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Research* **27**: 1939–1949.

Yardimci G, Ozadam H, Sauria MEG, Ursu O, Yan K-K, Yang T, Chakraborty A, Kaul A, Lajoie BR, Song F, et al. 2017. Measuring the reproducibility and quality of Hi-C data. *bioRxiv* 188755.

Zhan Y, Mariani L, Barozzi I, Schulz EG, Blüthgen N, Stadler M, Tiana G, Giorgetti L. 2017. Reciprocal insulation analysis of Hi-C data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes. *Genome Research* **27**: 479–490.

Zhang Y, Malone JH, Powell SK, Periwal V, Spana E, Macalpine DM, Oliver B. 2010. Expression in aneuploid Drosophila S2 cells. ed. P.B. Becker. *PLoS Biol* **8**: e1000320.

Zink D, Bornfleth H, Visser A, Cremer C, Cremer T. 1999. Organization of early and late replicating DNA in human chromosome territories. *Experimental Cell Research* **247**: 176–188.

Zorn C, Cremer C, Cremer T, Zimmer J. 1979. Unscheduled DNA synthesis after partial UV irradiation of the cell nucleus: Distribution in interphase and metaphase. *Experimental Cell Research* **124**: 111–119.

1.