Dissertation
submitted to the
Combined Faculties for the Natural Sciences and for Mathematics
of the Ruperto-Carola University of Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

Presented by
Tibor Pakozdi, B.Sc.
Born in Osijek, Croatia

Date of oral examination:
13/04/2015

Computational analyses of regulatory elements during embryonic development

Referees: Dr. John Marioni

Prof. Dr. Michael Boutros

Table of Contents

1. SUMMARY

1.1 Summary in English

All animals begin their life from a single initial cell. A cascade of cellular transitions and proliferation events during embryonic development eventually gives rise to tissues and organs. Their functionality is highly dependent on the composition of cellular machines - proteins. Since majority of cells contain same genetic material, the regulation of how this material is used, in other words how genes are expressed, results in a specific protein composition in cells that is critical for their individual function and thereby the proper development of metazoan organisms.

During my doctoral studies, I have focused my research efforts on understanding how gene expression is regulated by enhancers, non-coding elements that regulate spatiotemporal gene expression, thus acting as the decisive 'controllers' in the establishment of cellular identity during development. My interests revolve around two key topics in enhancer biology: How do specific transcription factors modulate enhancer activity, and what is their role in the formation of tissues such as mesoderm? What is the topological organization of enhancer-to-promoter interactions during development, and how do they relate to other regulatory features, such as the chromatin state?

To address these questions, I collaborated with experimentalists within the Furlong lab on three projects, each using embryonic development in *Drosophila melanogaster* as a model system: 1. Characterization of the recruitment of the repressive Polycomb complex (PhoRC) in a specific cell type during development; 2. Dissecting properties of enhancer to promoter interactions in two spatial (whole-embryo and mesoderm), and temporal (3-4h and 6-8h of development) contexts in the largest 4C-seq study up to date; 3. A description of more global chromatin topology, using high-resolution Hi-C data.

Using extensive statistical analyses, I found that a significant portion of PhoRC binding sites overlap with mesodermal developmental enhancers, have strong association with other Polycomb proteins, and cause repression of both enhancer and gene activity. The topological studies revealed that these enhancers tend to form a vast amount of long-range interactions, even within the compact *Drosophila* genome. These interactions remain largely unchanged, even though the embryo undergoes severe morphological transitions from multipotency to cellular specification during these stages, and correlate with the occupancy of paused RNA Polymerase II.

Taken together, these results have provided important new insights into how enhancer function, from the recruitment of transcription factors and their role in embryonic development, to the deciphering of the topological organization of chromatin in three-dimensional nuclear space.

1.2 Deutsche Zusammenfassung

Alle adulten Formen metazoischen Lebens beginnen ihr Dasein als einzelne Zelle. Durch eine Reihe von Zellteilungs- und Proliferationsereignissen bringt diese Zelle schließlich die verschiedenen Gewebe und Organe hervor. Mit fortschreitender Entwicklung differenzieren sich die Zellen innerhalb eines Organismus dabei zunehmend voneinander. Das Genom dieser Zellen ist jedoch, mit wenigen Ausnahmen, das Gleiche. Die Unterschiede liegen darin, wie und wann individuelle Gene aktiviert oder stillgelegt werden.

Im Zuge meiner Promotionsarbeit habe ich den Schwerpunkt meiner Forschung auf das Verständnis gelegt, wie Enhancer, nicht-kodierende DNA Elemente, welche die räumlich-zeitliche Genexpression regulieren, agieren, um die zellulare Identität während der Entwicklung festzulegen. Innerhalb der Enhancerbiologie konzentriert sich mein Interesse dabei auf zwei Schlüsselthemen: 1. Wie regulieren spezifische Transkriptionsfaktoren die Aktivität von Enhancern und welche Rolle spielen sie bei der Ausbildung von Geweben wie dem Mesoderm? 2. Wie sieht die topologische Organisation von Enhancer-Promoter-Interaktionen während der Entwicklung aus und in welchem Zusammenhang steht diese mit anderen regulatorischen Merkmalen wie dem Chromatinzustand?

Um diese Fragen zu beantworten habe ich mich an drei spezifischen Projekten beteiligt, von denen jedes auf das Verständnis eines Aspektes der Funktion von Enhancern zielt: 1. Die Charakterisierung von Bindungsstellen des reprimierenden, Polycomb-rekrutierenden Komplexes PhoRC während der frühen *Drosophila melanogaster* Embryogenese; 2. Die Entdeckung neuer genomischer Interaktionen zwischen distalen und Promoter-proximalen Enhancern in verschiedenen entwicklungsbiologischen Kontexten, als Teil der bisher größten 4C-seq Studie; 3. Das Verständnis der genomweiten Chromatintopologie in der Entwicklung mittels hochaufgelöster Hi-C Daten.

Unter Verwendung statistischer Analysen großer Datensätzen habe ich zeigen können, dass ein signifikanter Teil der PhoRC-Bindungsstellen mit im Mesoderm aktiven, entwicklungsspezifischen Enhancern überlappt, im engen Zusammenhang mit anderen Polycomb-Proteinen steht und die Reprimierung sowohl der Enhancer- als auch der Genaktivität verursacht. Meine topologischen Analysen haben ergeben, dass Enhancer dazu neigen eine große Anzahl von Interaktionen mit langer Reichweite innerhalb des kompakten *Drosophila*-Genoms zu formen.

Überraschenderweise waren, trotz der Tatsache, dass unsere Daten einen Zeitraum der Entwicklung abdecken, der von erheblichen morphologischen Veränderungen und Gewebespezifizierung gekennzeichnet ist, die identifizierten genomischen Interaktionen auffallend stabil und stimmten oft mit Bindungsstellen pausierender RNA-Polymerase II überein.

Zusammenfassend stellen die Ergebnisse meiner Forschung einen wichtigen Beitrag zu dem Fachgebiet der Genomik dar und zu unserem Verständnis von Enhancern, von dem spezifischen Kontext der Rekrutierung von Transkriptionsfaktoren und ihrer Rolle in der Embryonalentwicklung, bis hin zur Entschlüsselung der topologischen Organisation der DNA im dreidimensionalen Raum des Zellkernes.

2. INTRODUCTION

2.1 Regulation of gene expression

Higher eukaryotes, including humans, show great diversity in their morphology and in environments they inhabit. The underlying variation in complexity of different organisms is however not reflected in the difference in number of genes. For example, humans have ~22,000 protein-coding genes, while roundworm *Caenorhabditis elegans* has ~20,500, despite the fact that adult roundworm organism contains just 959 somatic cells and a simpler range of cognitive behavior (Kaletta, T. & Hengartner, M. O. 2006). It is now thought that this apparent discrepancy is reflected in the larger regulatory landscape of the human genome, rather than the increasing number of protein-coding genes (De Laat, W. & Duboule, D. 2013). There are multitude of regulatory points that altogether contribute to the identity of the cell, including control of gene expression, post-transcriptional regulation of messenger RNA (mRNA), translational control, and post-translational modifications of proteins. During my doctoral studies, I have been particularly interested in pre-transcriptional gene regulation, especially within the context of embryonic development. The genome is the core part of living cells that contains all hereditary and functional information encoded in the deoxyribonucleic acid (DNA). Although its sequence is composed from only the four nucleotides: adenine (A), thymine (T), guanine (G), and cytosine (C), it consists of reproducible and context-invariant signatures representing transcriptional start sites, start and stop codons, core promoters and non-coding elements (Ben-Tabou de-Leon, S. & Davidson, E. H. 2007), together forming a complex regulatory network that gives identity to each individual cell. In the following paragraphs, I will summarise the importance, functionality, and spatial topology of genomic elements that have a particular role in embryonic development.

2.2 *Cis*-regulatory modules

Stereotypic biological processes such as embryonic development require precise activation or repression of key developmental genes, which is accomplished through the integration of many regulatory inputs at regions of DNA named *cis*-regulatory modules (CRMs; Arnone, M. I. & Davidson, E. H. 1997; Wittkopp, P. J. & Kalay, G. 2012). These modules include the immediate upstream region next to the transcriptional start sites (TSS) - promoters, but also more distant, either intronic or intergenic DNA loci such as enhancers (increasing the levels of RNA production), insulators (associated with decrease of target gene expression) or silencers (completely preventing the onset of transcription). In the following paragraphs I will detail the sequence composition, chromatin context and function of such elements.

2.2.1 Promoters

Promoters are essential regulatory regions that play two key roles in the gene regulation: 1. They contain sequences required for the assembly of the core transcriptional machinery; 2. They integrate information from other *cis*-regulatory modules to control the final rate of transcriptional activity. Present upstream of the transcriptional start site of each gene, promoters are more strictly structured elements compared to other CRMs (in terms of their nucleotide composition), but also contain binding sites for transcription factors. TF binding causes the displacement of nucleosomes and statistical positioning both upstream of promoters and downstream of the TSS (Mavrich, T. N. et al. 2008).

In contrast to enhancers, much of the regulatory logic encoded by promoter DNA is relatively well understood. Core parts of the promoters are located between 10 and 50 base pairs upstream of the TSS and contain distinctive sequence signatures such as DRE, TATA-box, INR, and DPE motifs at which basic transcriptional machinery (including general transcription factors such as TFIIA/B/D and others) are being assembled before the process of transcription occurs (Ohler, U. et al. 2002). Also found within this pre-initiation complex (PIC) is Polymerase II (PolII), enzyme catalyzing the transcription of DNA into RNA. PolII occupancy is enriched within 200bp region downstream of TSS and its presence within the PIC is a prerequisite of transcription.

Apart from these distinctive motifs, promoters are also typically characterized by several epigenetic signatures, including the excess of CpG island methylation in humans and trimethylation of the lysine 4 of histone H3 (H3K4me3; Zhou, V. W. et al. 2011). Due to the binding of significant number of proteins, some of which directly interact with the DNA, promoters regions upstream of the TSS are also depleted in nucleosome occupancy. These nucleosome depleted regions (NDRs) are especially prominent upstream of highly active genes, and suggests a relationship between protein occupancy at the promoter and a genes' activity level (Jiang, C. & Pugh, B. F. 2009).

The second function of promoters is equally important. Within eukaryotes, genes are regulated via inputs from a diversity of signaling cascades and regulatory elements. These regulatory elements, some of which function to drive gene expression while others act to silence it, are found both proximal and distal to the regulated locus, and it is at the promoter that these diverse signals are integrated to direct appropriate gene expression in the correct biological context (Lenhard, B. et al. 2012). In short, promoters represent main regulatory sites that contain sequences for assembly of core transcriptional machinery and integrate regulatory information from other genomic elements.

2.2.2 Enhancers

Enhancers are short pieces of DNA (usually between 200bp-2kb in size) that positively regulate the rate of target gene transcription in order to control and fine-tune their spatiotemporal patterns of expression (Figure 1; Shlyueva, D. et al. 2014). Containing binding sites for transcription factors, enhancers can be located either upstream or downstream of the promoter region they act upon at varying distances albeit often within 50kb of target genes, preserving their function even when placed in different sequence context (De Laat, W. & Duboule, D. 2013; Shlyueva, D. et al. 2014). Regulation by enhancers is critical for some of the major biological processes such as embryonic development of metazoans, where precise activation of key genes within the regulatory network ensures the correct expression pattern and identity of cells within their spatiotemporal context. Often several enhancers act synergistically, which is well characterized phenomenon for *Drosophila melanogaster* genes such as Runt, which produces a seven-stripe expression pattern during embryogenesis, each of which can be linked to the individual enhancer element (Klingler, M. et al. 1996). Some of the enhancers (termed 'shadow') that act in groups can also partially overlap in their spatiotemporal activity, establishing robustness of the developmental programme against the changing environment (Hong, J.-W. et al. 2008). Deletion of enhancer sequences can lead to severe phenotypes. For example, a removal of MNE regions significantly reduces the levels of its target gene myelocytomatosis oncogene (*myc*) expression by 85% (Uslu, V. V. et al. 2014). A single point mutation in the ZRS enhancer, which is located around 1Mb away from the Sonic hedgehog gene (*shh*), causes ectopic expression of Shh in the developing limb bud, eventually causing severe deformations like polydactyly in humans and mice (Lettice, L. A. et al.

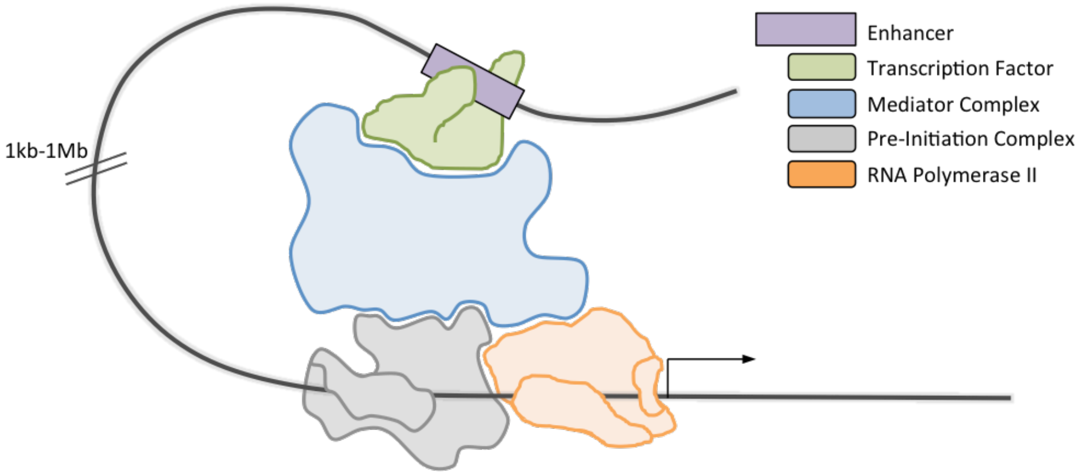

Figure 1. Transcriptional machinery assembled at gene promoter region. Distal regulatory regions, such as enhancers (purple box), can be located far from their target genes. They increase the expression of target gene by looping over in nuclear space in close spatial proximity to target promoter. Transcription factors (in green) bound to the enhancers interact with the mediator complex (in blue) that in turn regulates the gene expression by interacting with the pre-initiation complex and RNA Polymerase II.

2003; Sagai, T. et al. 2005; Visel, A. et al. 2009). Enhancers contain clear genomic signatures that can be used to determine their locations, including evolutionary conservation of sequence and binding motifs (Aerts, S. 2012), histone modifications (in particular H3K4me1; Creyghton, M. P. et al. 2010), and binding of transcription factors like P300 (Rada-Iglesias, A. et al. 2011). Consequentially, due to the frequent and direct occupancy of TFs, enhancer regions commonly contain displaced nucleosomes, and are thus readily detectable by the increase in cutting frequency from the enzyme such as endonuclease Deoxyribonuclease I (DNaseI; Crawford, G. E. et al. 2006), which prefer loci with open chromatin (including promoters). Alternative enzyme-free method for detecting enhancers is formaldehyde-assisted isolation of regulatory elements or FAIRE, which relies on the higher efficiency of cross-linking in nucleosome-wrapped DNA to extract and sequence the open chromatin (Giresi, P. G. et al. 2007). Although several different modes of enhancer regulation are possible (Villar, D. et al. 2014), one of the possible proposed mechanisms follow sequential steps: 1. Inactive enhancers are tightly wrapped around nucleosomes, which are obstructing the binding sites for the activating transcription factors; 2. Upon the binding of pioneer factor such as PU.1, nucleosomes are displaced and DNA becomes accessible to other proteins that can cooperatively bind to enhancer region (Barozzi, I.

et al. 2014); 3. Transcription factors bound to enhancers interact with the ~30-subunit mediator complex, which acts as a co-activator and mediates the interaction between the TFs on enhancers and promoter regions, releasing the RNA Polymerasee II from pre-initiation complex, thus activating the target gene (Ong, C.-T. & Corces, V. G. 2011). Spatiotemporal patterns of enhancers activity can be determined using imaging techniques. In short, a region of potential regulatory activity is placed in the construct upstream of minimal promoter and reporter gene, such as beta-galactosidase (lacZ). After generation of transgenic organisms, patterns of activity can be visualised using in situ hybridisation or luciferase assays at particular stages of development (Bier, E. et al. 1989; Manning, L. et al. 2012). An alternative is more high-throughput approach named STARR-Seq, which provides direct and quantitative measurement of enhancer activity by sequencing the self-transcribed enhancer products (Meyer, R. E. et al. 2013). Apart from the experimental methods, activity and expression pattern of enhancers can also be computationally predicted by either using the combination of epigenetic marks (H3K27ac, H3K79me3), and RNA Polymerase II (Bonn, S. et al. 2012), or co-occupancy of specific transcription factors (Zinzen, R. P. et al. 2009).

2.2.3 Insulators

Within the nuclear space, the genome can be very compacted, with many enhancers and genes overlapping, nesting, or looping over each other in ways that could drive gene expression if not carefully controlled for. This additional layer of control comes from proteins and DNA elements that act as insulators and insulating factors which act together to prevent unwanted transcriptional activities (Figure 2). The most famous example of an insulator binding protein in mammalian systems is CCCTC-binding factor (CTCF; Ong, C.-T. & Corces, V. G. 2014), which binds tens of thousands of loci genome-wide. Primarily associated with function as a barrier between domains of different activity (e.g, at the Wnt4 domain; Essafi, A. et al. 2011), large genomic experiments from ENCODE Consortium and others revealed a more complex role of CTCF. Apart from inhibiting gene transcription, CTCF was found to be co-binding with the architectural protein cohesin that is thought to play a major role in the establishment of intra-genomic interactions, including the stabilisation of enhancer-to-promoter contacts (Phillips-Cremins, J. E. & Corces, V. G. 2013). In *Drosophila melanogaster*, there exist several proteins likely to play an insulating function, such as ‘GAGA-factor, boundary-element-associated factor of 32 kDa' (BEAF-32), dCTCF, Centrosomal protein 190kD (CP190), and others, the binding of which has been shown to correlate with the demarcation of chromatin (into active and inactive domains) as well as with sites of chromosomal rearrangements suggesting that the insulator regions are evolutionary conserved between *Drosophilids* (Nègre, N. et al. 2010).

2.2.4 Silencers

Cell identity is established through the intricate interplay of regulatory networks, in which cell specificity is maintained not only though active gene expression, but also through negative regulation (repression; Beisel, C. & Paro, R. 2011). Transcriptional repression can be realized through several, non-exclusive mechanisms, including the condensation of DNA into heterochromatin (preventing the binding of activators; Chow, J. & Heard, E. 2009), positioning close to the nuclear membrane in lamina associated domains (LADs; Guelen, L. et al. 2008), the methylation of CpG islands (e.g., to prevent the activation of transposable elements; Bird, A. 2002; Cedar, H. & Bergman, Y. 2009), as well as direct regulation by binding of specific repressive factors such as Groucho, Brinker, and Polycomb group of proteins (PcG; Sparmann, A. & Van Lohuizen, M. 2006).

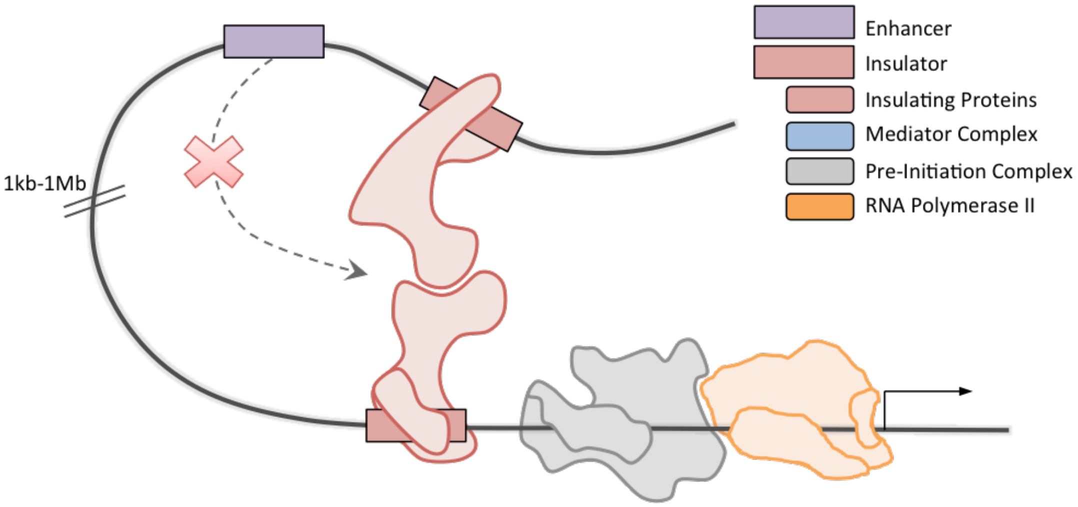

Figure 2. Insulating proteins blocking enhancer activity. Insulating proteins bound to the regulatory elements called insulators can form protein-protein interactions, thus remodeling the spatial organization of the genome. Different chromatin topology might prevent enhancers from reaching their target gene, thus blocking their activation.

2.3 Characterization of genomic functionality

2.3.1 Transcription factors

Transcription factors (TFs) are proteins that change the rate of gene transcription by forming protein-protein interactions with the pre-initiation complex and with RNA Polymerase II at the core promoter region upstream of the transcriptional start site. These interactions can either lead to an increase (activation), or a decrease (repression) in gene expression. Some transcription factors contain DNA-binding domains, and can thus directly recognize short (usually 6-12bp) stretches of DNA (known as sequence motifs), while others contain protein-binding domains and exert their function without direct binding to the DNA molecule. For example, the forkhead box D1 (FOXD1) protein contains a forkhead domain that recognises a 5'-GTAAACA-3' DNA motif and is required for kidney development (Fetting, J. L. et al. 2014), while melanocyte-specific gene 1 (MSG1) enhances the rate of Smad-mediated transcription by binding to P300/CBP coactivators without directly associating with DNA (Yahata, T. et al. 2000). Apart from the direct regulation of transcription, TFs are also involved in various other processes, including chromatin remodelling (especially so called 'pioneering factors', which displace nucleosomes to reveal binding sites to other factors, such as PU.1 and AP1; Zaret, K. S. & Carroll 2011; Biddie, S. C. et al. 2011), change the architecture of chromatin (e.g., 'straightening' of DNA bends by HMG1; Falvo, J. V. et al. 1995), and by aggregating on distal regulatory regions in a cooperative manner, where the overall of individual TF expression pattern plays an important role in determining the cell fate (Junion, G. et al. 2012). For the latter, there are three models as to how TFs come together to modulate the spatiotemporal activity of enhancers' target genes: 1. Enhanceosomes, in which a specific motif grammar (orientation and positioning of binding motifs) is required, alongside the presence of multitude of TFs to achieve activity (e.g., interferon-beta enhancer that consists of array of 8 different factors; Panne, D. et al. 2007); 2. The billboard model, where most, but not all TFs are required, but with fixed motif composition (Kulkarni, M. M. & Arnosti, D. N. 2003); 3. Collective binding, in which the overall mixture of activating TFs is necessary, but with both flexible sequence composition and variable DNA, and protein-protein binding scheme (Junion, G. et al. 2012). Overall, TFs act as essential parts in determining the activity of gene regulatory network, either through direct regulation of mRNA products, or by modulating the shape and accessibility of chromatin to other transcription-controlling factors at TSS proximal and distal regulatory sites.

2.3.2 Histone modifications

At its most basic level, chromatin organization consists of 147bp of DNA wrapped around nucleosomes separated by linker regions (Jiang, C. & Pugh, B. F. 2009). Nucleosomes are composed out of histone proteins H2A, H2B, H3, and H4, along with the linker histone H1. Each one of these histone subunits contains intrinsically disordered tails that protrude out of the nucleosome, and can be in direct contact with DNA (Peng, Z. et al. 2012). Containing many lysine, arginine and serine residues, histone tails can be subjected to post-translational modifications that can both influence and reflect the functional state of the DNA region that is wrapped around them (Zhou, V. W. et al. 2011). For example, a recent publication (Pengelly, A. R. et al. 2013) showed that a point mutation that replaced the 27th residue of histone H3 from a lysine to an arginine is sufficient to reproduce the mutant Polycomb phenotype, suggesting that the tri-methylation of K27 is essential for Polycomb repression, while acetylation was suggested as not playing a significant role apart from the potential antagonisation of repression (Pengelly, A. R. et al. 2013). Apart from methylation (mono-, di-, and tri-) and acetylation, histones can also undergo ubiquitination (e.g., H2AK119Ub, which is deposited by the PRC1 complex; Hu, H. et al. 2012), phosphorylation (such as on the H2AS129; Rossetto, D. et al. 2012), cronylation (H2BK5Kcr; Tan, M. et al. 2011), SUMOylation (H3-SUMO; Shiio, Y. & Eisenman, R. N. 2003; Nathan, D. et al. 2003), and others. Recent work has identified 67 previously unknown modifications, with an even larger variety possible, if not likely (Tan, M. et al. 2011). Histone modifications have been associated with range of different biological activities and features: 1. Active gene expression (e.g. H3K79me3); 2. Transcriptional repression (e.g. H3K27me3 and H4-SUMO; Shiio, Y. & Eisenman, R. N. 2003); 3. Exon-intron usage (H3K36me3; Kolasinska-Zwierz, P. et al. 2009); 4. Large-scale repression (heterochromatin markers H3K9me2/3; Muramatsu, D. et al. 2013); 5. Promoter location (H3K4me3); 6. Enhancer location (H3K4me1); 7. Enhancer activity (H3K27ac and H3K79me3; Bonn, S. et al. 2012). Since histone modifications provide such useful information and can be readily measured using genome-wide sequencing experiments like ChIP-Seq, integrating many different modifications within the same computational framework (e.g. Hidden

Markov Models; Ernst, J. & Kellis, M. 2010) continues to provide useful insight into the functional organisation of the chromatin.

2.4 Regulation of gene expression by Polycomb group proteins

Spatiotemporal regulation of gene expression is established and maintained by transcription factors, which can act as both activators and repressors of transcription. One of the latter system and key regulators of embryonic development are the evolutionarily conserved Polycomb Group (PcG) proteins (reviewed in Simon, J. A. & Kingston, R. E. 2009; Di Croce, L. & Helin, K. 2013). Since their initial discovery in *Drosophila melanogaster* as repressors of homeotic genes during the specification of the embryonic antero-posterior axis (Lewis, E. B. 1978), PcG proteins have since additionally been implicated in the regulation of many other biological processes, including mammalian development (Pietersen, A. M. & Van Lohuizen, M. 2008), the regulation of lineage factors (Bracken, A. P. & Helin, K. 2009), and cancer formation (Simon, J. A. & Lange, C. A. 2008). In the following paragraphs I will describe the different protein complexes formed by the PcG proteins, their functional roles in animals and plants, proposed mechanisms of repression and the features of the DNA sequences to which PcG are recruited in *Drosophila*.

2.4.1 PcG protein complexes

Molecular studies pioneered in *Drosophila* revealed that PcG proteins assemble into at least five distinct protein complexes (reviewed in detail in Bantignies, F. & Cavalli, G. 2011 and Beisel, C. & Paro, R. 2011): Pleiohomeotic repressive complex (PhoRC), Polycomb repressive complex 1 (PRC1), Polycomb repressive complex 2 (PRC2), dRing-associated factors (dRAF), and Polycomb repressive deubiquitinase (PR-DUB). PRC1 consists of four different proteins: Polycomb (Pc), Polyhomeotic (Ph), Posterior sex comb (Psc) and Sex combs extra (Sce or dRing). PRC2 is composed of Enhancer of zeste (E(z)), Extra sex combs (Esc), Suppressor of zeste 12 (Su(z)12), and Nucleosome-remodeling factor 55 (Nurf-55). PRC2 contains a histone methyltransferase activity encoded in the E(z) subunit and deposits H3K27me3 – a mark that is subsequently recognised by PRC1 (Beisel, C. & Paro, R. 2011). PcG protein complexes are recruited to Polycomb response elements (PREs), nucleosome-depleted regions in the vicinity of PcG target genes, by Pleiohomeotic (Pho) and Scm-related gene containing four MBT domains (Sfmbt), which together form the Pho-repressive complex (PhoRC). Apart from H3K27me3, PcG proteins also deposit mono-ubiquitin on lysine 119 of histone H2A (H2AK119Ub). This process is catalysed by both PRC1 and dRAF – a complex that is compositionally related to PRC1 and contains Psc, dRing and the Lysine (K)-specific demethylase 2 (dKdm2)). Indeed, the E3 ligase activity shared by PRC1 and dRAF is encoded in the Sce/dRing subunit, and H2AK119 monoubiquitination has been proposed to be involved in the inhibition of RNA PolII elongation (Stock, J. K. et al. 2007; Zhou, W. et al. 2008). Lastly, H2AK119Ub is removed by the PR-DUB complex, composed of Calypso and Additional sex combs (Asx). The importance and mechanism of action of monoubiquitination of H2A119 still remains to be clarified, since despite being opposite enzymatic reactions, removal of the PRC1-catalyzed H2AK119 monoubiquitination by PR-DUB leads to gene silencing as well (Scheuermann, J. C. et al. 2010). PcG protein complexes are also evolutionary conserved in mammals, however many more complex variants are present in mammals because several mammalian orthologs of each *Drosophila* PcG protein arose from gene duplication (Bantignies, F. & Cavalli, G. 2011). For example, *Drosophila* Pc has five paralogous CBX genes (Di Croce, L. & Helin, K. 2013). In total, compared to 15 *Drosophila*, there are 37 mammalian Polycomb-related genes. This leads to an even greater possible variety of mammalian PcG complexes, such as in the case of PRC1 where 180 different complexes are theoretically possible (Di Croce, L. & Helin, K. 2013).

2.4.2 Regulation of developmental genes and cell fate, and role in cancer formation

Polycomb-based regulation of gene transcription has been implicated in a number of biological processes, including: 1. Silencing of developmental genes in embryonic stem cells, such as Oct4, Sox2 and Nanog that are poised for activation before differentiation occurs (Lee, T. I. et al. 2006); 2. Imaginal disc formation and control of cell cycle, with evidence that PcG proteins directly regulate key cell cycle genes such as CycB, and cause misesxpression of other developmental regulator genes like *eve*, *Doc2/3* and *tsh* when mutated, leading to aberrant tissue formation (Oktaba, K. et al. 2008); 3. X-chromosome inactivation (PcG components Ring1B and Mel18 were shown to be recruited to the inactive X chromosome and are responsible for H2A119UB deposition; de Napoles, M. et al. 2004); 4. Maintenance of epigenetic memory, where PcG proteins and their antagonistic activators – the so-called

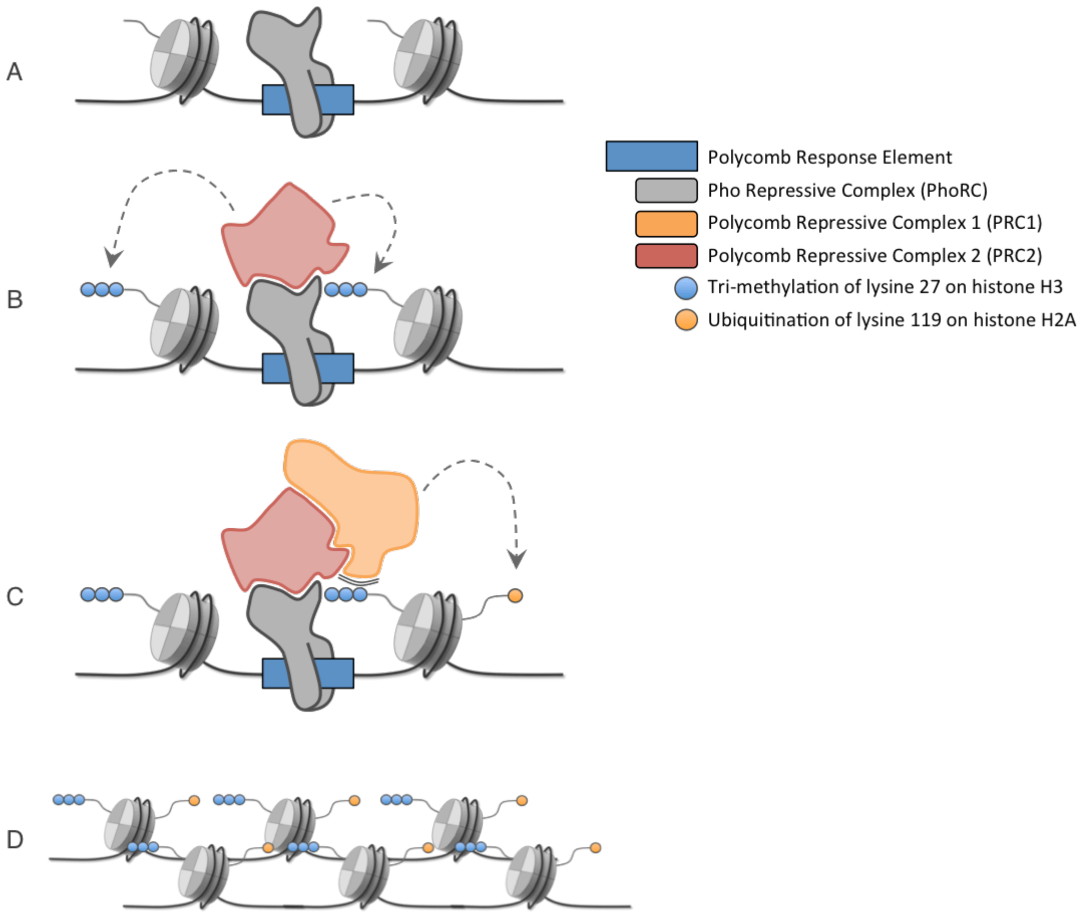
A
B
C
D
Polycomb Response Element
Pho Repressive Complex (PhoRC)
Polycomb Repressive Complex 1 (PRC1)
Polycomb Repressive Complex 2 (PRC2)
Tri-methylation of lysine 27 on histone H3
Ubiquitination of lysine 119 on histone H2A

Figure 3. Mechanism of PcG repression. The intial model of PcG-based repression starts with the binding of Polycomb recruiting complex PhoRC to PREs (A). Proteins from PRC2 group form protein-protein contacts with PhoRC, and deposit repressive histone mark H3K27me3 (B). This histone mark is then recognized by chromodomain of Pc from PRC1 group (C). Another post-translation modification H2AK119Ub is placed by enzymes from PRC1 group leading to the overall compaction of chromatin, which might result in formation of larger nuclear structures - Polycomb bodies (D).

Trithorax group (TrxG) proteins - maintain the appropriate gene expression throughout the cell cycle, possibly through the involvement of non-coding RNAs and bound proteins (Schmitt, S. et al. 2005; Ringrose, L. & Paro, R. 2007); 5. Regulation of lineage factors. Studies in murine system that progresses from stem cells to neural progenitors and terminal pyramidal neurons revealed that many progenitor-specific genes are targets of dynamic Polycomb regulation (Mohn, F. et al. 2008); 6. Cancer formation through the deregulation of cell fate transcription factors that leads to accumulation of cells lacking the ability to differentiate, as in the case of aberrant recruitment of PcG proteins by PLZF-retinoic acid receptor-alpha fusion protein leading to acute promyleocytic leukaemia (Villa, R. et al. 2007).

2.4.3 Mechanism of PcG repression

An initial model of *Drosophila* PcG protein recruitment to and silencing of target genes was the following (Figure 3; Schwartz, Y. B. et al. 2006; Schuettengruber, B. et al. 2007; Simon, J. A. & Kingston, R. E. 2009; Bantignies, F. & Cavalli, G. 2011; Beisel, C. & Paro, R. 2011): regulatory regions (PREs) are directly bound by the sequence-specific DNA binding protein Pho, which is part of PhoRC complex. This complex then recruits members of PRC2 group, which contains E(z) that (catalyzes the deposition of repressive histone mark H3K27me3, which was recently shown to be a crucial substrate for PRC2; Pengelly, A. R. et al. 2013). The chromodomain of the Pc subunit of the PRC1 complex recognizes H3K27me3, thus leading to PRC1 recruitment. PRC1 then places another repressive mark H2AK119Ub, leading to wide-spread repression outside of PREs and compression of local spatial context into nuclear Polycomb bodies, which are enriched in high concentration of PcG proteins that prevents the ATP-dependent nucleosome remodeling by SWI/SNF complex and assembly of pre-initiation complex (Sparmann, A. & Van Lohuizen, M. 2006; Pirrotta, V. & Li, H.-B. A 2012). This sequence of events has recently been challenged, since it was shown that PRC1 can interact with PhoRC, independently of PRC2 recruitment (Schoeftner, S. et al. 2006; Tavares, L. et al. 2012). Exact mechanisms within different biological contexts still remain unresolved.

2.4.4 Polycomb response elements (PREs)

Polycomb proteins are recruited to the regulatory regions named Polycomb response elements (PREs). As noted in the paragraph 2.4.1, PcG complex is mainly recruited through PhoRC, the Pho subunit of which contains a zinc-finger binding domain and a clear DNA binding motif 'GCCAT' (Oktaba, K. et al. 2008). Other recruiters such as Dorsal switch protein 1 (Dsp1), GAGA-factor (GAF), and others have been implicated in the potential recruitment of PcG complexes (Müller, J. & Kassis, J. A. 2006), but unlike Pho, they did not show a clear Polycomb phenotype upon mutation (Schuettengruber, B. et al. 2007). In *Drosophila* there are currently dozens of defined PREs that have been experimentally validated using two distinct genetic tests for their capacity to repress the transcription of a linked reporter gene in transgenic *Drosophila* (Poux, S. et al. 2001). More specifically, these PREs were tested for:: 1. Repression of the bxd enhancer that drives the expression of Ubx promoter, fused to the lacZ reporter gene; 2. Repression of a downstream mini-white gene, which causes the white-eye adult phenotype upon gene silencing. Unlike PcG proteins themselves, PREs are not evolutionary conserved, since only two PREs have been identified in mammals so far (Beisel, C. & Paro, R. 2011), including the PRE-kr that regulates the MafB gene and contains highly a conserved hcPRE segment, and is bound by PRC1 and PRC2, albeit at varying affinity (Sing, A. et al. 2009). The other identified mammalian PRE is a 1.8kb region named D11.12 that is located between HOXD11 and HOXD12 genes (Woo, C. J. et al. 2010).

2.5 Spatial organization of the genome

The spatial organization of the genome plays an important role in regulation of core biological processes, including maintenance of cell identity during embryonic development of metazoan organisms. Since regulatory elements and factors that occupy them often reside far away from their target genes (in terms of linear genomic distance), the genome has to adopt a certain conformation in order to bring them into close spatial proximity. There are several levels of chromatin organization that all together contribute to the final rate of transcriptional activity: 1. Primary biochemical properties of the DNA, such as winding of the strands, handedness, and interaction with the different variants of the histone proteins that form nucleosomes, basic building blocks of the chromatin; 2. Chromatin compaction, where highly condensed regions (heterochromatin) generally suppress gene expression, while more relaxed parts are in general accessible to the binding of proteins like transcription factors (euchromatin); 3. Physical interactions between different regions of the genome, such as those that occur between regulatory elements and their target genes; 4. Chromosome territories and positioning of chromosome domains within the nucleus, such as the parts of chromatin that contact the nuclear lamina, which are known to have lower levels of gene activity. In the next few paragraphs I will detail different structures and recently developed methods that allowed us to probe the topological organization of the genome from wrapping around nucleosomes to the large-scale domains and nuclear compartments.

2.5.1 Nuclear architecture

With only a few exceptions, each cell within the metazoan organism contains largely the same genetic information (bar differences in somatic SNPs), and yet reaches different functionality depending on which genes it expresses. Transcriptional regulation is influenced by mechanisms that act at many different levels of genome organization, including the relative position within the nucleus. The nucleus is an organelle contained within a double membrane, the inside surface of which is covered with a structured lattice consisting of laminar proteins that support its structural shape within the cell (Aebi, U. et al. 1986). Apart from mechanical support, nuclear lamina can also be in contact with particular parts of the genome, forming distinct domains (LADs; Guelen, L. et al. 2008). Genes contacting the nuclear lamina generally show lower levels of expression and gene density, and tethering a gene to the inner nuclear membrane (in this case the lac operator; Finlan, L. E. et al. 2008) is sufficient to silence the gene's expression. A reciprocal experiment of repositioning a gene away from the nuclear lamina to the interior resulted in activation of gene expression (Therizols, P. et al. 2014).

The remaining portion of the genome that is located away from the nuclear membrane is also not randomly organized, demonstrated by both computational and experimental methods. In mammals, chromosomes themselves occupy distinct spatial domains termed chromosome territories (Cremer, T. & Cremer, M. 2010). Computational simulations of polymer dynamics, with the parameters inferred from chromosome conformation capture experiments, suggested that the folding is not in equilibrium over the whole space, but rather forms a fractal globule structure (Lieberman-Aiden, E. et al. 2009). Imaging of the global nuclear structure showed that particular classes of proteins of known biological function are not uniformly distributed throughout the nucleoplasm, but are rather enriched at particular nuclear bodies. These bodies contain a high concentration of proteins that perform a particular function, for example proteosomic degradation of proteins in clastosomes (Lafarga, M. et al. 2002), or production and modification of RNA molecules in cajal bodies (Morris, G. E. 2008). Some of the proteins found in nuclear bodies are also involved in the regulation of gene expression, either activating (when surrounded by the increased density of RNA Polymerase in transcription factories, which also process the mRNA molecules; Rieder, D. et al. 2012), or silencing the transcription (through association with repressive Polycomb bodies; Pirrotta, V. & Li, H.-B. A 2012). To summarise, on the nuclear scale, DNA is non-randomly organised into high-order structures that define the further likelihood of genomic interactions, and also bring genes into proximity of molecules that can directly modulate their expression.

2.5.2 Topologically associated domains (TADs)

Chromosome territories and nuclear bodies represent higher-level chromatin organization that can correspond to large differences in gene activity. Recent advances in genome-wide 3C-based technologies such as Hi-C and 5C have however revealed another, more fine-scale topological structures of the genome that are characterized by a higher frequency of local interactions, separated by sharp boundaries between individual domains (Dixon, J. R. et al. 2012; Gibcus, J. H. & Dekker, J. 2013). These structures, named topologically associated domains (TADs), are spread throughout the genome and range in size from the megabase scale in mammals, to an average of hundred kilobases in genomic distance in *Drosophila*. Increased spatial proximity of inter-TAD contacts was also confirmed by imaging experiments. For example, distance between probes measured by DNA-FISH in mouse embryonic stem cells was significantly shorter when placed within same domains, than between neighboring domains defined by 5C experiment (Nora, E. P. et al. 2012).

There are at least four different ways in which genomically distant loci might come into spatial proximity and structure topological domains: 1. Specifically through direct protein-protein interactions (for example by the binding of insulator proteins such as CTCF and cohesin; Splinter, E. et al. 2006); 2. As bystander regions next to the crosslinked proteins; 3. Through random intertwining (DNA-DNA contacts) that is not stabilized through third party; 4. By colocalizing within the larger compartments, such as nuclear lamina or transcription factories (Dekker, J. et al. 2013). There seems to be a strong tendency of TADs to correlate with the underlying functional organization and activity of genes as supported by two lines of evidence: 1. High overlap with the boundaries of histone modifications representing different chromatin states (Sexton, T. et al. 2012); 2. Correspondence with the large regulatory domains containing enhancers exhibiting similar pattern of activity (Symmons, O. et al. 2014). TADs were also found to be stable regulatory units of replication timing (Pope, B. D. et al. 2014), although their organisation is probably only restricted to the interphase part of the cell cycle, since chromosome conformation capture experiments on mitotic chromosomes suggest the existence of a more homogeneous folding state (Naumova, N. et al. 2013).

Since TAD boundaries seem to be particularly important in structuring interactions within the genome, several studies have looked for signatures that might underlie the regions between neighboring domains. Within the spanning genomic distance of several kilobases, TAD boundaries have been shown to correlate with the increased occupancy of insulating factors such as CTCF, enrichment of housekeeping genes compared to genome-wide levels, and increased occurrence of repetitive elements (Dixon, J. R. et al. 2012; Hou, C. et al. 2012; Gómez-Díaz, E. & Corces, V. G. 2014). However, the structural importance and functional implications of boundary elements still remains to be completely elucidated, as demonstrated by some cases such as the 58kb deletion of the region between Xist and Tsix topological domains containing CTCF sites, which did not lead to a complete merge of the neighboring domains nor significant reorganization of genomic contacts (Nora, E. P. et al. 2012). As an additional confirmation of their significance and in agreement with functional organization of genomic elements, TADs were also shown to be evolutionary conserved between humans and mice (Gorkin, D. U. et al. 2014; Dixon, J. R. et al. 2012). Despite some examples of differences in chromatin architecture corresponding to the progressive activation as in Hox genes during embryonic development of mouse (Noordermeer, D. et al. 2014; Williamson, I. et al. 2014), TADs were found to be largely invariant between different cell types (Dixon, J. R. et al. 2012) and during the differentiation from embryonic stem cells to neural progenitors and fibroblasts (Nora, E. P. et al. 2012), suggesting that while genomic contacts within the topological domains can differ, overall structure of the genome seems to be preserved. In short, TADs represent one of the basic structural units of the genome that significantly correlates with the underlying functional signature, and seems to be both evolutionary conserved and largely invariant to the different spatiotemporal contexts.

2.5.3 Local compaction

Apart from the partitioning of genome into more accessible euchromatin, and compacted heterechromatin, the DNA molecule is wrapped around nucleosome particles, the regulation and composition of which can influence the binding of transcription factors, the initiation and processivity of transcription and genome replication (Jiang, C. & Pugh, B 2009). Nucleosomes neutralize the high negative charge of DNA, and present the basic scaffold for further structuring into the 30nm fiber and whole-chromosome folding. The composition of nucleosomes varies according to the activity of the genes it contains. For example, highly expressed genes exhibit a statistically well-positioned nucleosome immediately downstream of their transcription start site (TSS), and a nucleosome depleted region just upstream of TSS (Yuan, G.-C. et al. 2005), presumably due to the binding of core transcriptional machinery and transcription factors (Mavrich, T. N. et al. 2008). There are two modifications of nucleosome that influence the rate of transcription: 1. Post-translational modifications of amino acid residues on the histone tail (the functional consequences of which are detailed in the section 2.3.2); 2. Variation in the composition of the core histones, with more dynamic and unstable variants H2A.Z and H3.3 replacing the canonical H2A and H3 around the highly active sites (Malik, H. S. & Henikoff, S. 2003). In short, the higher-scale interactions presented in the previous paragraphs seem likely to depend on the rigidity and composition of chromatin, which in turn is defined by the local organisation of the DNA molecule around nucleosomes.

2.5.4 Methods for probing the organization of genome

2.5.4.1 Chromosome conformation capture (3C)-based technology

Since the association of enhancers to their nearest genes as their potential targets is a very poor predictor of regulatory contacts (only 7% of such cases were observed in the recent 5C study; Sanyal, A. et al. 2012), mapping of genomic interactions genome-wide is essential for understanding the regulatory landscape (De Laat, W. & Duboule, D. 2013). Recently, chromosome conformation capture (3C) technologies enabled genome-wide discovery of specific genomic interactions (Dekker, J. et al. 2002). These methods are based on the idea that spatial organisation (with a significant proportion mediated by proteins) is stabilized by crosslinking, followed by shearing or enzymatic digestion of DNA and proximity ligation of fragments that were in contact. Summarised over the population of cells, the average interaction frequency can then be associated between different genomic elements. Depending on the further experimental design, researchers can focus on one-to-one (3C), one-to-many (4C), many-to-many (5C) or all-to-all (Hi-C) contacts, which represent different trade-offs of specificity and sensitivity of the assessed interactome.

2.5.4.1.1 3C

Developed as the first conformation technology (Dekker, J. et al. 2002), 3C uses specific primers on both the anchoring and target restriction fragment regions to determine the contact frequency using quantitative PCR amplification (De Wit, E. &

de Laat 2012). Although requiring rigorous controls before the interpretation of the data, 3C research demonstrated the occurrence of at least three types of genomic interactions that set the foundation for more high-throughput methods: 1. Looping of enhancer elements to their target promoters (e.g., the locus control region and beta-globin gene promoter; Tolhuis, B. et al. 2002), and discovery of novel enhancers (Gheldof, N. et al. 2010); 2. Confirmation that insulator proteins play an important role in establishment of interactions (Splinter, E. et al. 2006); 3. Existence of contacts between starts and ends of genes, which might facilitate transcription through RNA Polymerase II re-loading (O'Sullivan, J. M. et al. 2004).

2.5.4.1.2 4C

Instead of PCR amplification following the digestion with the first restriction enzyme, 4C technology uses a second restriction enzyme to form smaller fragments that result in circular products after the ligation (Simonis, M. et al. 2006). Using specific primers and inverse PCR, only products with the specific fragment ('viewpoint') are selected and sequenced (Splinter, E. et al. 2011). Unlike 3C, 4C measures all genomic interactions from viewpoints to other regions that were in spatial proximity (both intrachromosomal and interchromosomal). Since interactions can originate due to different events, including non-specific contacts (see 2.5.2.1.1), an extensive statistical analysis of 4C data is required to precisely determine the significantly interacting fragments (van de Werken, H. J. G. et al. 2012). An increased number of newly discovered interactions were used together with external datasets to try to decipher whether spatial organisation of the genome is a cause or consequence of gene transcription. Although some specific cases such as beta-globin gene do show the emergence of new interactions upon activation in a specific tissue (Simonis, M. et al. 2006), most of the spatial structure seems to be preserved and unaltered, even upon the strong external stimuli (Hakim, O. et al. 2011). Such examples are showing how high-throughput assessment of spatial interactions can help to understand the regulatory genome, despite the potential constraints of the technology (e.g. known enhancer elements in the short distance from target promoters might not be decipherable due to limits in resolution; De Wit, E. & de Laat 2012).

2.5.4.1.3 5C

If a mixture of restriction fragment end based oligos is used instead of a single primer, a conformation of a whole region of interest can be evaluated using 5C many-to-many method (Dostie, J. et al. 2006). Having an increased throughput of the 3C technology, 5C contact matrices that are built based on the frequency of local genomic interactions provide an important insight into the topological organisations of the chromatin besides individual contacts, and can be used to model the phenomenon such as the openness of chromatin (Baù, D. et al. 2011). Although genome conformations are now most often interpreted from whole-genome Hi-C matrices, 5C method still provides useful information into local chromatin structure without having to compromise on sensitivity and specificity in measuring the local topological organisations.

2.5.4.1.4 Hi-C

Hi-C is the first method to expand 3C-based technology to the measurement of all chromosome interactions along the entire genome (Lieberman-Aiden, E. et al. 2009). Utilising the unbiased selection of biotin-labeled fragments, Hi-C has been used to describe the structure of chromatin on the larger scale (i.e. TADs and genomic compartments) compared to previous 3C methods in a number of systems, including human cell lines (Dixon, J. R. et al. 2012; Lieberman-Aiden, E. et al. 2009), mouse embryonic stem cells (Dixon, J. R. et al. 2012), *Drosophila* embryos (Sexton, T. et al. 2012), yeast (Duan, Z. et al. 2010), and even single cells (Nagano, T. et al. 2013). Although seemingly the most promising 3C-technology, there are several concerns to be considered when interpreting Hi-C results: 1. Compromised specificity, due to the large sample space and lack of sequence coverage; 2. Lower resolution (~1Mb interaction blocks in human genome; Lieberman-Aiden, E. et al. 2009). Recent reports suggest that both of these difficulties have been somewhat alleviated through the improvement in sequencing technology (increasing the read coverage, usage of more frequent enzymatic cutters, and improved statistical analysis with rigorous filtering; Jin, F. et al. 2013), and methodology (by performing proximity ligation within the cell nucleus, increasing the human Hi-C data to kilobase resolution, which is two orders of magnitude higher than previously reported; Rao, S. S. P. et al. 2014).

2.5.4.2 ChIA-PET

All of the methods that were mentioned so far result in quantification of genomic interactions independent of the nature of their origin, including capturing a lot of non-specific contacts. Since large portions of the spatial organisation are based on the protein-protein contacts, a method was developed that involves immunoprecipitation of an interaction protein of interest, named ChIA-PET (Chromatin Interaction Analysis by Paired-End Tag Sequencing; Fullwood, M. J. et al. 2009). Using specific antibodies, ChIA-PET is able to isolate events where a DNA-DNA contact is mediated by the same protein. Recent studies involved analysis of two protein-based interactions, one with the estrogen receptor alpha (Fullwood, M. J. et al. 2009), and CTCF (Handoko, L. et al. 2011). Although the prospect of isolating specific protein-based interactions seems very intriguing, the method has low signal-to-noise ratio and captures only a small subset of the total interactome (De Wit, E. & de Laat 2012).

2.5.4.3 Imaging techniques

Despite the fact that 3C-based technologies greatly contributed to our understanding of spatial organisation of DNA within the nucleus, they have a unifying flaw (besides single-cell Hi-C; Nagano, T. et al. 2013) that the genomic interactions are measured over the population of cells, which only represents the average topology, and might miss the transient or cell-specific contacts. For that reason, a complementary method based on the cell imaging needs to be used to confirm and agree with the spatial proximity inferred from 3C methods.

2.5.4.3.1 Fluorescence in situ hybridisation (FISH)

The most common technique for measuring the spatial distance between two genomic loci within the nucleus is fluorescence in situ hybridisation (FISH), which utilises the hybridisation of specific fluorescently-labelled probes to the DNA locus of interest. Nuclear structures such as chromosome territories and aggregations of gene-rich regions towards the centre of nucleus has been visualised with FISH and confocal imaging in both fixed and living cells (Edelmann, P. et al. 2001; Müller, I. et al. 2010). Although the FISH method is low-throughput, depends on the probe design and resolution limits of microscopy, a recent example of discrepancy between 5C topological map and the discovered spatial organisation by the FISH method on the HoxD locus in a PRC1 mutant demonstrates the need for imaging as the crucial complement in evaluation of genomic interactions (Williamson, I. et al. 2014). The main concern of measuring spatial proximity only through 3C-based interactions is that the chromatin composition (availability of residues for cross-linking), alongside the partial decondensation prior to the ligation step in 3C methodology may result in unspecific contacts that are not support by the imaging results (Williamson, I. et al. 2014). Such phenomenon was demonstrated on the example of interaction between beta-globin gene, and it's enhancer where only 1% of actual contacts were subjected to proximity ligation (Gavrilov, A. A. et al. 2013). For this reasons, and due the constant advancement of microscopic techniques for fixed samples like STED-FISH (Zhang, W. I. et al. 2014) and 3D-SIM (Cerase, A. et al. 2014), imaging remains an indispensable complement to genome-wide detection of genomic interactions.

3. Methods

3.1 Occupancy of PhoRC on developmental enhancers

3.1.1 Sample preparation, sequencing and alignment

To study repressive regulation of genes within the context of development of mesodermal tissue in *Drosophila melanogaster,* I collaborated with Jelena Erceg from the Furlong group who used a modified ChIP-Seq protocol (Bonn, S. et al. 2012) named BiTS-ChIP-Seq to sort Twist-expressing cells using Fluorescence-Activated Cell Sorting (FACS) technology. Cells were sorted from transgenic embryos that expressed a green flourescent protein (GFP) under the promoter of the specific mesodermal transcription factor Twist. Following the sorting and enrichment of specific DNA fragments bound by the Polycomb recruiters Pho and dSfmbt, samples were sequenced on either Illumina GA-IIx (Pho) or Hi-Seq machines (dSfmbt; Minoche, A. E. et al. 2011). To make biological replicates of dSfmbt from the Hi-Seq sequencing more comparable to Pho, reads in FASTQ files were trimmed to 36bp - length of the sequenced Pho reads used in our study. Reads were aligned to the *Drosophila melanogaster* genome version 3 (July 2006; Celniker, S. E. et al. 2003) using BWA v0.7.5a (Li, H. & Durbin, R. 2009), allowing for two mismatches and no gaps (-n 2 -o 0). Additionally '-I' parameter was used for Pho samples that contained Phred+64 quality encoding. Only non-duplicate uniquely aligned reads with the 'XT:A:U' tag were kept for further analysis. Reads aligned to unassembled contigs (U/Uextra) and the mitochondrial genome (M) were discarded.

3.1.2 Estimation of the fragment length

True protein binding loci are known to be located between the pile-ups of reads on forward and reverse strands, since the protected region that was crosslinked with DNA is not captured by short single-strand sequencing (Park, P. J. 2009), requiring the shift or extension of each read towards the 3' direction. Shift sizes were estimated by selecting maximum Pearson's correlation coefficient value between all of the + and - strands reads mapped on the chromosomal arm 2L (Figure 5). Exact values of each estimate is summarised alongside the read counts in Table 1. For visualisation and further analysis, I merged biological replicates into single alignment files for each developmental stage and sample combination using samtools v0.1.19-44428 (Li H. et al. 2009).

3.1.3 Peak calling

cisGenome v2.0 (Ji, H. et al. 2008) was used to locate the enriched ChIP regions from two biological replicates compared to the 4-6h and 6-8h control (input) replicates using default parameters, with the exception of extending shifted reads by 36bp (-e 36), setting a higher neighbouring peak threshold (-maxgap 200), and defining a stringent standardised t-statistic cutoff (-c 3.5). A union of Pho peaks at two different developmental stages was taken to remove redundancy, followed by intersection with dSfmbt peaks to define the PhoRC loci. Flybase annotation v5.9 (St Pierre, S. E. et al. 2013) was used throughout the analysis in this study. Read counts after each filtering step from biological replicates are summarised in Table 1.

3.1.4 Normalization and visualization

Difference in sequencing depth between the libraries was corrected by using Reads Per Genome Coverage (RPGC) normalization (Bonn, S. et al. 2012), in which the total read count coverage was multiplied by the ratio of read length (36bp) and mappable genome size (1.35e+08). Corrected coverage was summarised into 20bp bins. For visualisation tracks, ChIP samples were additionally subtracted with the input control.

3.1.5 Defining the list of developmental enhancers

Three previously published datasets containing the information on regulatory regions were used to construct the list of developmental enhancers: 465 literature-based CRM activity database 2 (CAD2; Bonn, S. et al. 2012), 8008 mesodermal enhancer based on the binding of 5 transcription factors (Zinzen, R. P. et al. 2009), and 4041 Tinman-bound cardiac enhancers (Junion, G. et al. 2012). Since some of the transcription factors used to define the regions were same between the different sources, several steps were taken to remove the redundancy between the datasets: 8008 enhancers that overlapped with CAD2 enhancers were removed, together with the cardiac enhancers that overlapped with TF8008 set, resulting in the unique set of 9,513 developmental enhancers. Since the focus of my research was on the distant regulatory regions, an additional TSS proximity filter within the distance of 500bp was applied (6,606 remaining). Finally, regions that overlapped with 6-8h H3K4me3 peaks were removed to avoid the potential confounding effect of unannotated TSS regions, creating the final set of 5,949 enhancers.

3.1.6 Construction of the background regions

To evaluate the significance of Pho colocalization on the defined set of developmental enhancers, a background set of regions was constructed by randomly sampling 124,800 starting positions over the *Drosophila melanogaster* genome, followed by the calculation of mappability (defined as percentage of mapped reads per base pair), local GC content, region width, chromatin accessibility (defined as number of RPGC-normalized input reads) and TSS distance for both observed (1,248 peaks) and expected regions. A sampling algorithm from the R package MatchIt was used (Ho, D. et al. 2011) with mahalanobis distance to find the equal number of expected regions which most closely matched in the mentioned genomic properties to the observed set (Figure 4). Significance of enhancer occupancy by Pho was then calculated using Fisher's Exact Test.

3.1.7 Motif discovery

Discovery of *de novo* motifs was performed on the feature-separated *Drosophila melanogaster* genome version 3 sequences defined 100bp +/- around the Pho peak summit using MEME v.4.9.1.1 (Bailey, T. L. et al. 2009), with the following parameters: '-dna -oc promoter -nostatus -maxsize 1000000 -mod zoops -nmotifs 20 -minw 5 -maxw 50 -revcomp seq.fa’.

3.1.8 Analysis of the vertebrate homologue YY1

A list of human enhancers based on the DNaseI hypersensitivity signal (DHS) from five different cell lines (K562, IMR90, HELA, H1 and GM12878) and YY1 peaks were downloaded from the ENCODE consortium (Rosenbloom, K. R. et al. 2013; Kent, W. J. et al. 2010). Since only the positions of the highest DHS enrichments were reported, enhancers were extended 1kb upstream and downstream of peak summits. To calculate the likelihood of observing YY1 on human enhancers, a background set of YY1 peaks was constructed by matching the genomic properties such as GC content (50bp resolution), mappability (50bp resolution), proximity to the TSS (human genome annotation GRCh37p10; Flicek, P. et al. 2014; Kasprzyk, A. 2011), and width of the observed regions as previously described in paragraph 3.1.6. Rank-ordered Z-scores of YY1 signal on developmental enhancers (all three classes, filtered by the TSS proximity of 2kb; Rada-Iglesias, A. et al. 2011), together with the histone modifications (H3K4me1, H3K4me3, H3K27me3) were based on the human embryonic stem cell data from ENCODE (Rosenbloom, K. R. et al. 2013).

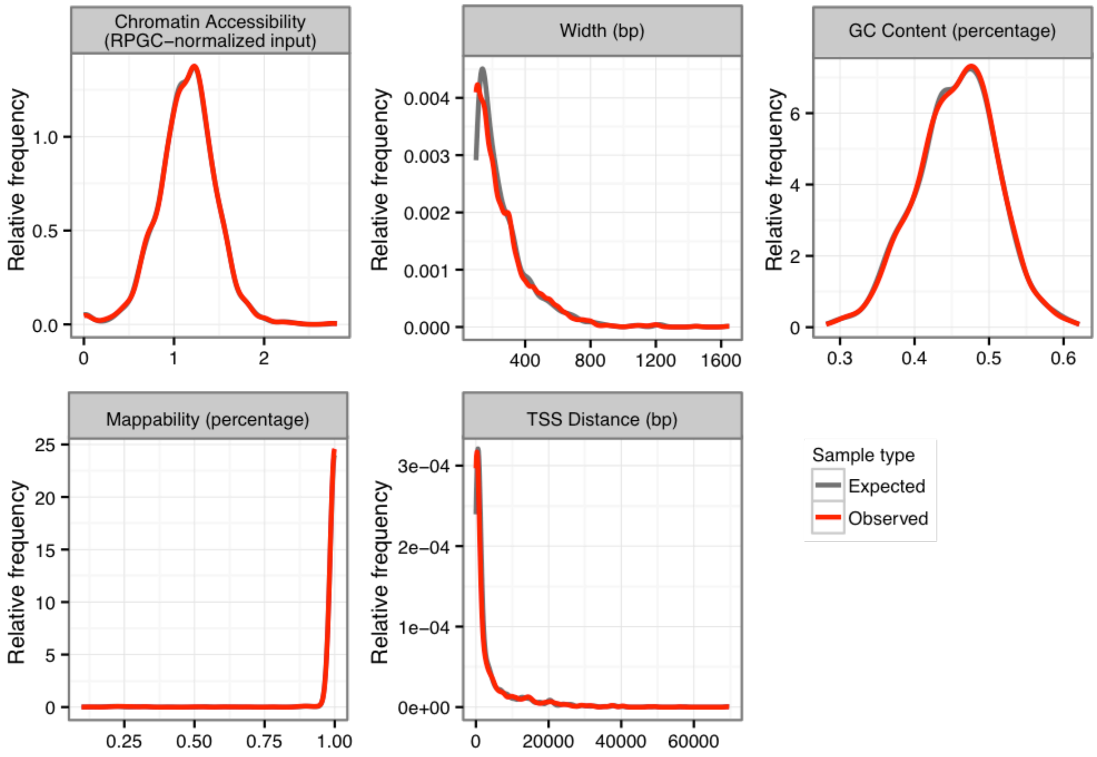

Figure 4. Background regions for the PhoRC peaks. Regions were found through matching of 5 genomic properties: chromatin accessibility, width, GC content, mappability and TSS proximity. Read distributions show the observed signals on the 994 genome-wide PhoRC loci, while grey represent the matched set of equal size (see Methods).

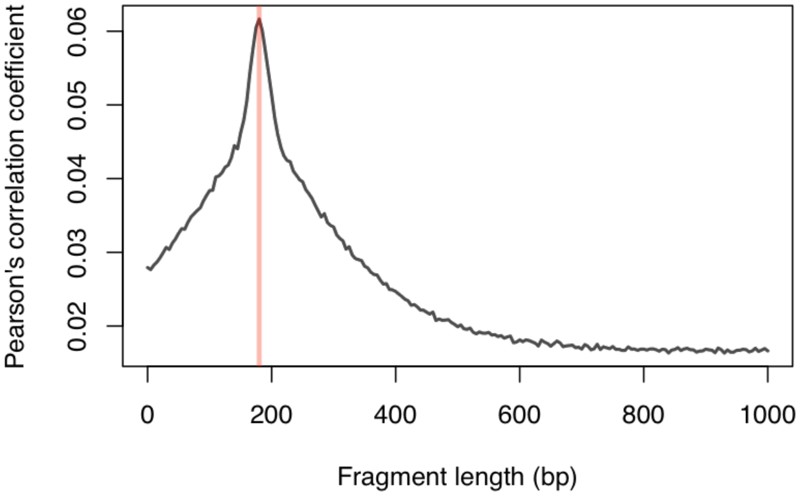

Figure 5. Estimation of the optimal fragment shift length. Represented using the first biological replicate of Pho at 6-8h, based on the cross-correlation of forward and reverse strand reads. The grey curve represents the correlation coefficient corresponding to the shift on the x-axis, while the vertical red bar shows the highest point of correlation, corresponding to the optimal fragment length of 180bp.

3.2 4C-seq interactions in embryonic development

3.2.1 Circularized chromosome conformation capture followed by sequencing (4C-seq)

Genomic interactions were determined using a variant of the chromosome conformation capture technology named 4C-seq on *Drosophila* embryos, as previously described (van de Werken, H. J. G. et al. 2012; Ghavi-Helm, Y. et al. 2014). Briefly, chromatin in proximity was crosslinked using formaldehyde, followed by digestion with a first restriction enzyme DpnII. After reverse crosslinking and digestion with a second restriction enzyme NlaIII, fragments were circularized and amplified using inverse PCR reaction with viewpoint-specific primers. Samples were prepared for high-throughput sequencing using standard protocols. All of the wet-lab part of the experiment was performed by Furlong lab postdoc Yad Ghavi-Helm.

3.2.2 Alignment and estimation of expected counts

Sequenced reads were demultiplexed and aligned to the *Drosophila melanogaster* genome version 3 (July 2006; Celniker, S. E. et al. 2003) using the Novoalign algorithm with default parameters. Aligned reads were associated with the '4C reference genome' based on the DpnII-defined fragments as described (Ghavi-Helm, Y. et al. 2014). To estimate the significance of observing an interaction, variance stabilised read counts were fit using a monotonous local regression function per each viewpoint and experimental condition. Z-scores were calculated from the residuals of the fit, and converted to P-values using a standard Normal distribution. Alignment, pre-processing, normalization of data, and estimation of the fit was performed by a Felix Klein (Huber group).

3.2.3 Determination of significant fragments

Fragments were set as significant if the Z-score value was higher than 3 (nominal P-value of 0.001) in both biological replicates while having less than 10% false-discovery values in at least one of the replicates. Since neighboring region are, of necessity, closely located in 3D space, significant fragments within 1kb of each other were iteratively merged until no other significant fragment was found within the set limit. A unique set of interactions was made from the non-redundant union of significant regions from all 4 experimental conditions (whole-embryo at 3-4h and 6-8h, and mesodermal tissue at 3-4h and 6-8h) with the precedence of regions defined at whole-embryo 6-8h in case of overlap.

3.2.4 Differential analysis of 4C interactions

Differential interactions per viewpoint between either spatial (whole-embryo over mesoderm-specific condition), or temporal (6-8h over 3-4h) contexts were found using by fitting the experimental variance to a negative binomial using DESeq2. To compare the conditions, values of the local fit from each fragment were used as normalization vectors. Fragments were called differentially significant if the difference between conditions well within a 10% false-discovery rate and more than an absolute Log2-fold change in normalized read-counts between conditions.

3.2.5 Association of interactions with genomic elements

Significant interactions were associated with genomic annotations of developmental enhancers derived from data on histone modifications, transcription factor occupancy, and the literature as described in Methods 3.1. Other genomic loci were taken from the FlyBase genomic annotation v.5.47 (St Pierre, S. E. et al. 2013), with promoters being defined as regions +/- 1kb around each isoform transcription start site. In case where interacting region overlapped both promoters and enhancers, it was assigned to the category with which it had a higher percentage overlap.

3.2.6 Construction of background interactions

Background interactions were computationally selected from the larger set of genome-wide DpnII fragments, while matching for the genomic properties of the significant 4C interactions such as GC content, mappability, and width of interacting regions using Mahalonobis distance from the MatchIt package (Ho, D. et al. 2011).

3.2.7 Definition of active and non-active genes

Categorization of genes into 'Active' and 'Non-Active' groups was based on RNA-Seq read counts (Figure 6; Graveley, B. R. et al. 2011) from all biological replicates at 2-4h and 6-8h of embryonic development. Since the overall distribution of gene expression has a bimodal shape, thresholds for each replicate were determined based on the local minima of each distribution, which was previously shown to be a good estimator of gene activity (Hebenstreit, D. et al. 2011). To be labeled as 'Active', transcript counts for each gene had to be either above the threshold for both biol. repl. at 2-4h, or in 4 out of 5 for 6-8h.

3.2.8 Topological and syntenic domains

For analysis of interactions containment (placement within the same regions as the viewpoint) within topological domains (defined in Sexton, T. et al. 2012), background set of comparison was found by sampling the distance from log-normal distribution, estimated specifically for each viewpoint and condition. Observed conservation (positioning within the same domain) of significant interaction was then compared to randomly placed background set. Similar approach was done for syntenic blocks, defined previously using whole-genome alignments between *D. melanogaster* and 4 other species (*D. ananassae, D. pseudoobscura, D. mojavensis* and *D. virilis*;

Engström, P. G. et al. 2007). Comparison between observed and expected conservation was done using a two-sided Mann-Whitney U Test.

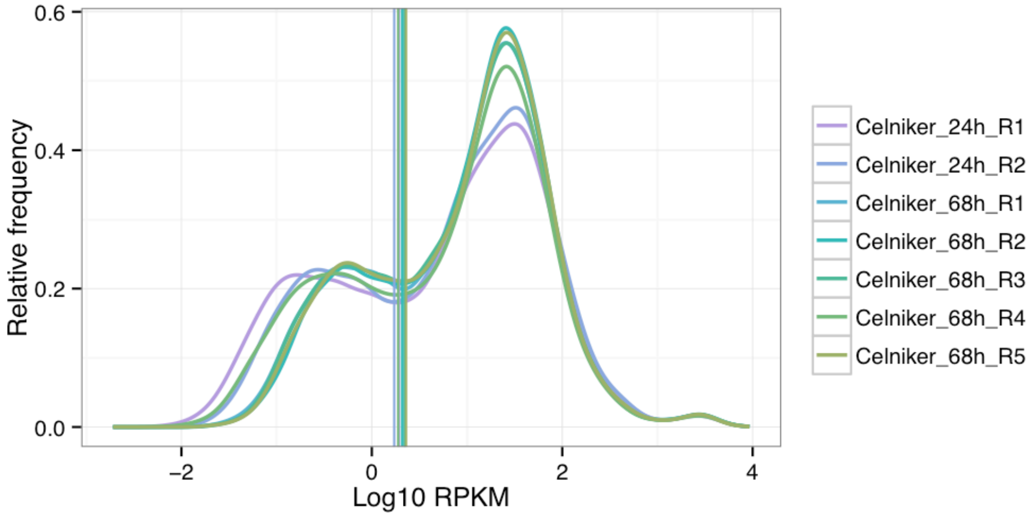

Figure 6. Distribution of read counts from whole-embryo RNA-Seq experiment. From two stages of embryogenesis - 2-4h, and 6-8h. Vertical lines indicate the replicate-specific thresholds that are positioned at the local minima of the distribution, splitting the genes into ‘Non-Active’, and ‘Active’ categories.

3.3 High-resolution chromatin interactions from Hi-C

3.3.1 Experimental setup and alignment of Hi-C reads

Hi-C experiment was performed as previously described (Sexton, T. et al. 2012). Briefly, DNA regions in close spatial proximity were crosslinked with formaldehyde, followed by treatment with restriction enzyme DpnII, which recognizes and cuts DNA at specific GATC sequence motif. Smaller fragments were subjected to proximity ligation, forming the circular products that were sonicated and prepared for sequencing by Yad Ghavi-Helm, a collaborator in Furlong Group. Samples were sequenced with Hi-Seq technology resulting in 104bp, paired-end reads (Minoche, A. E. et al. 2011), which I mapped with BWA-mem alignment algorithm (from the BWA 0.7.5a version; Li H. 2013) to the *Drosophila melanogaster* genome version 3 (July 2006; Celniker, S. E. et al. 2003) using default parameters. Following the alignment, I removed all reads associated with mitochondrial (M) and unassembled chromosomes (U, Uextra), and used the information about mapping quality in the aligned file to distinguish between multiply-mapped and uniquely-mapped read pairs (0 for former, and all other for latter). After the filtration of uniquely mapped reads, I split the products of several DpnII fragments ligations from the dataset by search for BWA-mem –specific flag 'SA:Z' (representing chimeric reads, which were part of the same sequenced read, but contain fragments within that were aligned to different locations in the genome). Lastly, I removed all duplicate read pairs with SAMtools v.0.1.19-44428cd (Li H. et al. 2009), and compressed files into bam format. For the purpose of this analysis, I considered only *cis* reads (intrachromosomal).

3.3.2 Filtering of *cis*-pairs and calculation of biases

The following methods for estimating the expected read score frequency and P-values using Negative-Binomial model is based on the modified computational protocol for human cell line high-resolution Hi-C study from Bing Ren's laboratory (Jin, F. et al. 2013). After alignment and preprocessing, I distinguished between three different types of read pair orientations by different SAM flags in the ensuing manner: samestrand (65, 129, 113, 177), outward (81, 83, 145, 147), and inward (97, 99, 161, 163), due to the previously observed frequency bias for short-distance pairs (Jin, F. et al. 2013). I labeled each read pair with the corresponding orientation and associated the 5'-most position with the overlapping DpnII fragment. I inferred the interaction length from insert size, where I kept only positive instances (since SAM format reports both pairs). To calculate the potential read frequency bias, I used fragment-level information (summarised in the 1kb bins up to 50kb) of outward and inward read pairs, compared to samestrand ones for short insert size distances. There exists a clear increase in the bias for fragments within a close genomic range, which reaches the expected 50% frequency level at 10kb length (Figure 7). I use the following information to remove all orientation-based biases for inward and outward reads when constructing the anchor-interaction matrix.

3.3.3 Filling contact matrices with Hi-C reads

To construct the count matrix, I divided the whole genome into DpnII fragments based on the recognition site 'GATC', which resulted in a total of 356,504 fragments. For each fragment, I calculated GC content (frequency of G and C nucleotides), mappability (average number of 104bp reads mapping to each fragment such that each base pair is completely covered), and width. Due to the large matrix size (for example, a single chromosomal arm 'chr2L' contains 60,628 fragments), I used the HDF5 format to store the large data point structures (e.g. 60,628 x 60,628 anchor-interaction points for chr2L). Each read-pair was then added to the pre-constructed 64bit matrix using the 10kb orientation-bias filter, and followed by the addition of a transposed matrix to maintain the symmetry. These HDF5 matrices were then used for all subsequent processing and analyses.

3.3.4 Estimation of the expected score

To find the significantly interacting regions I calculated the expectation score, which represents the background level of Hi-C reads as a result of biases in genomic properties like mappability, GC-content, and fragment-length. I used the modified method that was previously described in the high-resolution Hi-C article on human cell lines (Jin, F. et al. 2013). In short, to estimate the expectation value, for each anchor-fragment contact, I performed a search for all other anchor-interaction instances that match in either fragment length and anchor-to-fragment distance bin (along with min. requirement of 20% mappability), over the whole chromosome for L-bias correction (distance bias), or GC content within the space of 400kb from anchor midpoint (F-bias). Both biases are computationally intensive for estimation considering the large number of possible anchor interactions within the same DpnII-generated intra-chromosomal fragment space. To simplify the estimation, continuous

GC-content and fragment width values were binned into 20 equally-sized groups, while fragment to anchor distance length was sorted in 2kb bins (up to 400kb, which was set as a *cis*-detection limit). Read counts for the whole parameter space (20 GC-content x 20 width x 201 distance = 80,400 values) were summarised using the arithmetic mean. Expectation scores were then calculated for all anchors within the Dm3 genome (each row of the DpnII-based matrix) by multiplying the L-bias, F-bias and mappability per each anchor-fragment interaction.

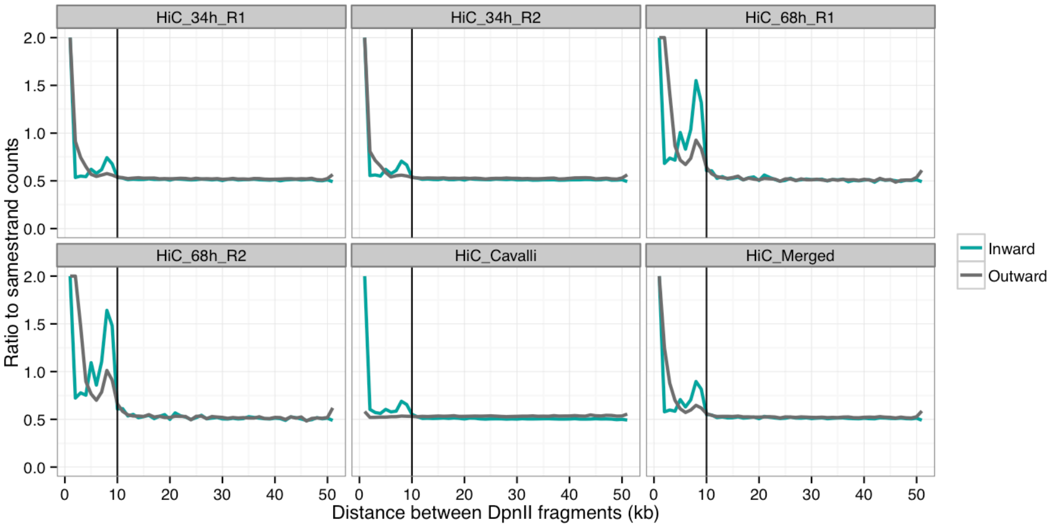

Figure 7. Bias in read count frequency for inward and outward orientated read pairs. Ratio against the same strand pairs levels at 50% at 10kb distance between fragments, which was set as filtering threshold. Merged sample is the sum of all technical and biological replicates used in the study.

3.3.5 Calculation of P-values

When average expectation scores are plotted against the variance of the observed read counts in the range of 0.95 x K to 1.05 x K, where K is the increasing expectation values from 1 to 10, I observed the previously described linear relationship between the two variables (Jin, F. et al. 2013). Due to this unique property, calculation of the P-value from the integer-based two-parameter negative binomial model is greatly simplified to the real number of failures that equals (expected score) / (beta - 1), and success probability = (1 / beta), where beta is the slope of fitted linear regression line for the previously described relationship. Since the model does not explicitly correct for the known proximity bias (lower p-value for short distance interactions; Thongjuea, S. et al. 2013; Williams, R. L. et al. 2014) that contrasts the limits of Hi-C technology, I penalised all anchor-fragment contacts within 10kb (where the read count bias occurs) by 10% higher expectation score than estimated for every 250bp, such that e.g. at 9kb distance expectation score is 140% higher, while at 5kb distance it increases to 300% (Figure 9). All P-values for interactions within 1kb of their anchors were set to 1. Beta (slope of the linear regression) is determined separately for each chromosome, but tends to be very similar across the estimations (Figure 8).

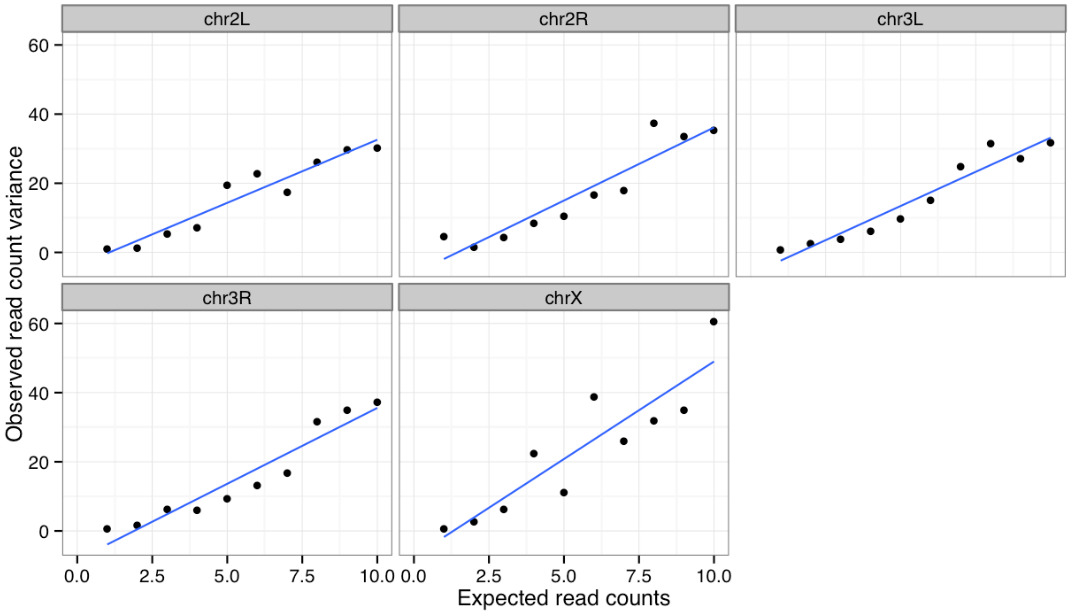

Figure 8. Linear relationship between the estimated expectation and observed read counts. Slope of the fitted line is used in chromosome-specific calculation of the P-values from the simplified Negative-Binomial model.

3.3.6 Informative P-value cutoff for Hi-C fragments

In order to set the meaningful and informative P-value threshold for defining the anchor-fragment interactions, I used previously evaluated 4C-seq viewpoints, which are comparable to the Hi-C data due to the usage of the same restriction enzyme DpnII (Figure 10). First, I extracted all of the rows of the Hi-C matrix corresponding to the designed 4C-seq viewpoint DpnII fragments and their corresponding non-zero interacting fragments. Following the association of 4C and Hi-C dataset, I used the 1,983 whole-embryo 6-8h interactions as a positive, and related negative, training set from the background matching process in 4C-seq study. Using the wide range of p-value threshold, and the described training set, I calculated the empirical sensitivity as ((#true positives) / (#true positives + #false negatives)), and specificity as ((#true negatives) / (#true negatives + #false positives)). Only the fragments that passed both

5% P-value and had at least 10 observed read counts thresholds were then eligible for a merge test, where all the significant interactions that are within the 100bp of each other were joined, such that the minimal P-value is kept and the boundaries redefined to the edges of the fragments in vicinity, until no other significant fragment is found within 100bp.

3.3.7 Association with genomic features

Significant interactions were associated with genomic features defined by the FlyBase genome annotation v.5.47 (St Pierre, S. E. et al. 2013) and list of 5,057 developmental enhancers (see 3.1 Methods). For cases in which a fragment overlaped with multiple features, a choice between enhancers and promoters was made depending on which feature had the higher percentage of overlap (as in 3.2 Methods).

3.3.8 Ontological analyses of Hi-C interactions

I used Ontologizer (Robinson, P. N. et al. 2004), a collection of methods for calculating the overrepresentation of the gene ontology (GO) terms associated to the given gene list. I tested the interacting genes for their enrichment in either biological process, localization or function with the 5.47-based files for gene ontology analysis, using the following line: 'java -jar Ontologizer.jar --go gene_ontology.obo --association slim_association.fb --population background.txt --studyset target.txt --mtc Bonferonni-Holm --calculation Parent-Child-Union’. Complete mRNA-producing gene set was used as a background reference.

3.3.9 Differential Hi-C analyses

Differential contact frequency analysis was performed using DESeq2 (Love, M. I. et al. 2014), with full quantile normalisation (EDAseq; Risso, D. et al. 2011), local fit of the dispersion estimates, P-value calculation using Negative-Binomial Wald test with default parameters, and independent filtering (Figure 11), where I estimated that multiple testing correction was effective for fragments with at least 10 counts.

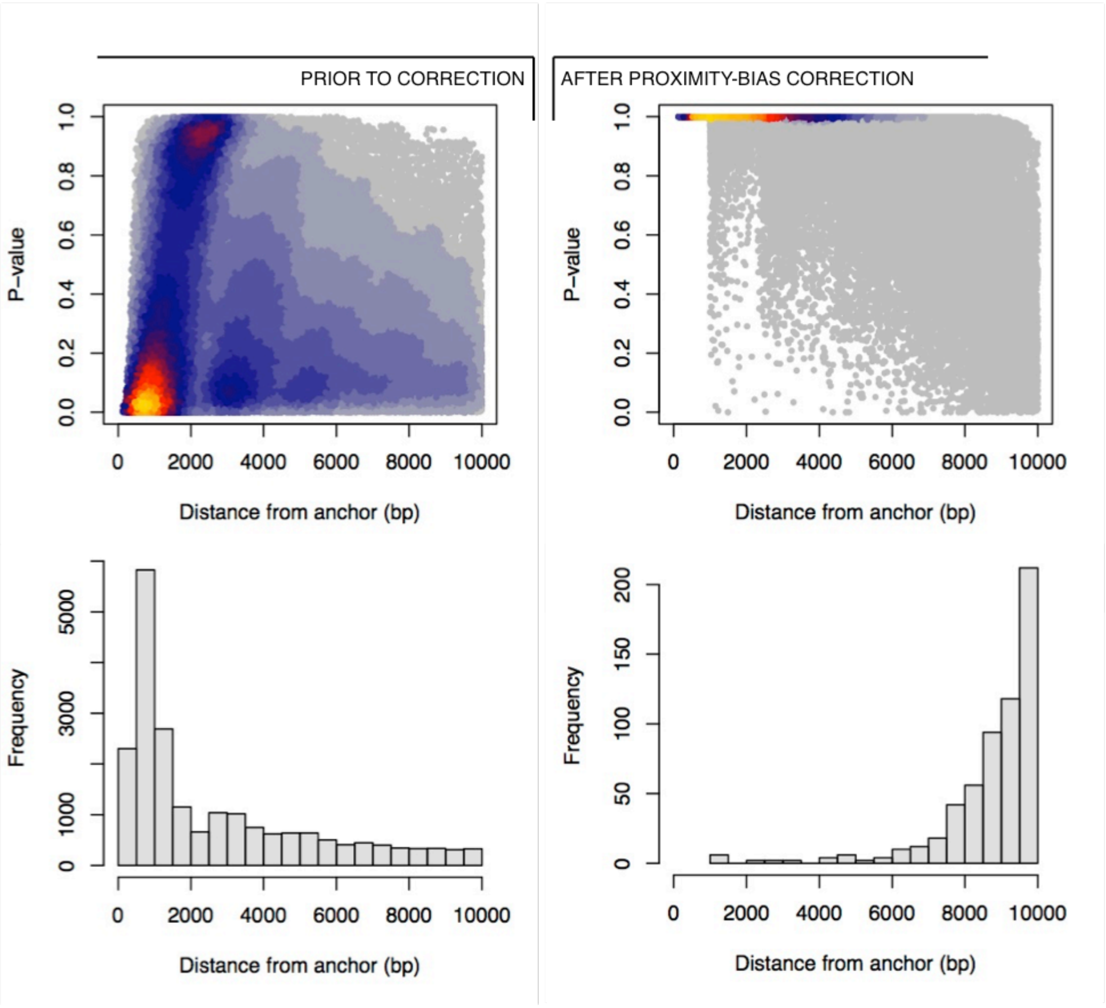

Figure 9. Anchor proximity bias correction. Left side of the figure represents the high number of significant fragments within 2kb of their respective anchors. After the application of 10% increase of expectation score for each 250bp closer to anchor location, and a strict 1kb cut, significant interactions start resembling the expected distribution considering the technical limitations of Hi-C technology.

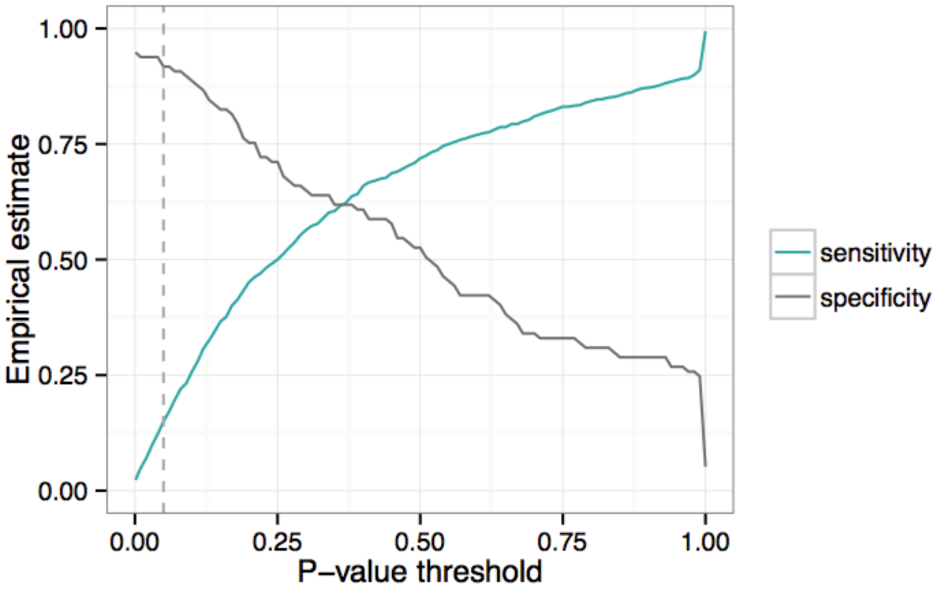

Figure 10. Informative P-value threshold selection for Hi-C fragments. Based on the sensitivity and specificity of 4C-seq positive (WE 6-8h), and negative interaction sets. Dashed line represents the chosen threshold of 5%, where specificity of 90.72%, and sensitivity of 22.01% is observed.

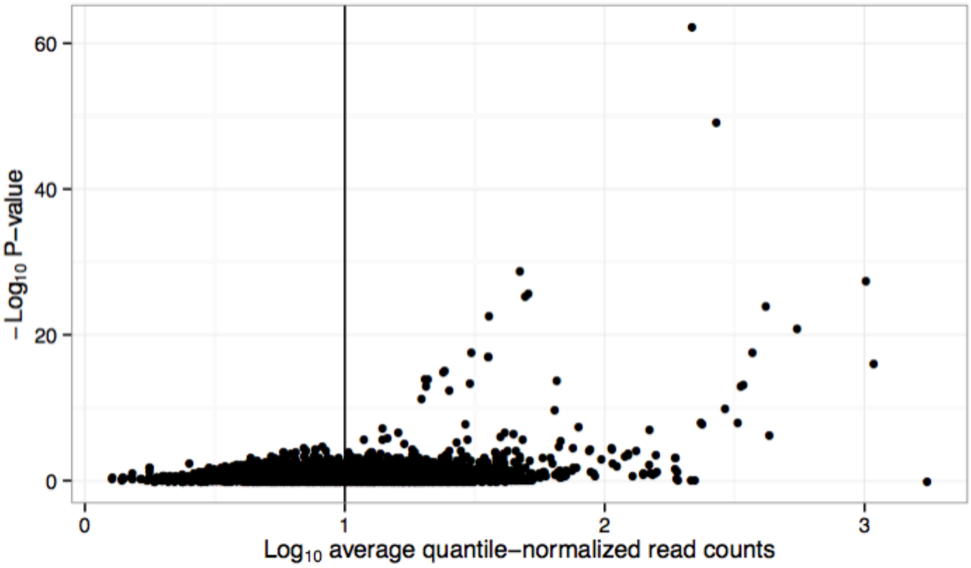

Figure 11. Independent filtering analysis for differential Hi-C contacts. Multiple testing correction with Benjamini-Hochberg method was performed only for fragments with an average read count higher than 10 (points right of the vertical line).

4. RESULTS

4.1 The role of the Polycomb complex at developmental enhancers

4.1.1 Description of the occupancy sites of the Polycomb recruiting complex PhoRC

4.1.1.1 Aims of the study

The correct development of metazoan organisms requires precise spatio-temporal regulation of gene expression and maintenance of cell identity, which is achieved through tight control imparted from regulatory elements such as enhancers through the action of transcription factors (TFs). TFs can either enhance (increase), or repress (decrease) the levels of gene transcription. To understand how repressive regulation operates within the context of mesodermal development, we used *Drosophila melanogaster* as a model organism and characterised the binding sites of two recruiters of the repressive Polycomb (PcG) system: Pleiohomeotic (Pho), and Scm-related gene containing four mbt domains (dSfmbt). All of the experimental work in this project, including the isolation of mesodermal nuclei and the preparation of samples for ChIP-Sequencing was performed by a collaborator in the Furlong lab, Jelena Erceg.

4.1.1.2 Quality assessment of tissue-specific Pho and dSfmbt ChIP-Seq data

I analyzed the data from chromatin immunoprecipitation followed by massively parallel DNA sequencing (ChIP-Seq) to identify occupied loci of two Polycomb recruitment proteins: Pho and dSfmbt that together form a DNA-binding complex called the Pho Repressive Complex (PhoRC), that is thought to in turn recruit the Polycomb Repressive Complexes 1 and 2 (PRC1 and 2) that effectuate gene silencing (reviewed in e.g. Simon and Kingston, 2013). Experiments were performed in either a mesoderm-specific (Pho, dSfmbt) or whole-embryo spatial context (dSfmbt), and at 4-6h (Pho) or 6-8h (Pho and dSfmbt) of embryonic development, corresponding to changes from multipotency of cells (stages 8-9) to their specification into muscle primordia (stages 10-11). Following the alignment, pre-processing and normalization, I have examined the occupancy of both TFs on known Polycomb Response Elements (PREs). A clear increase in ChIP signal compared to the input levels was observed on all known sites of Polycomb binding, including the *bxd* and *iab-2*_(1.7) PREs, P{C4-826} and others. Recovering these known sites, along with the high reproducibility (Figure 12 and 13), demonstrates the high quality and sensitivity of our dataset.

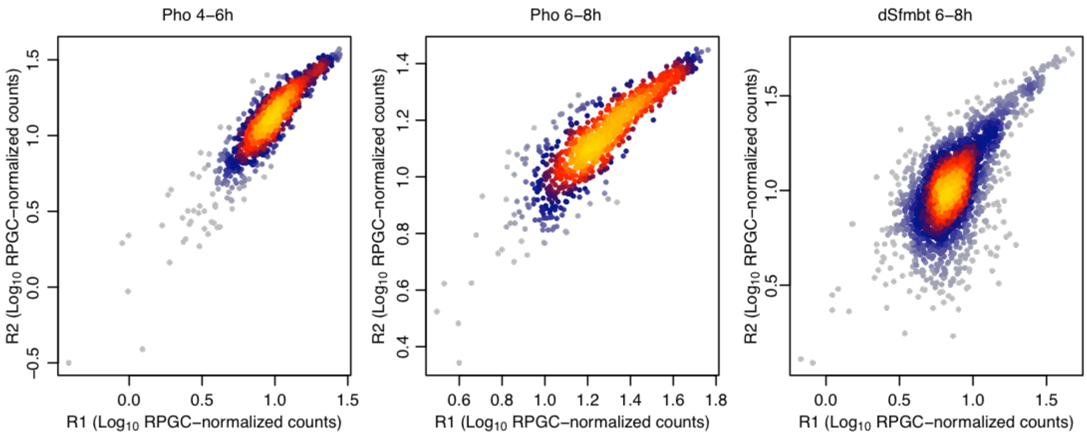

Figure 12. High reproducibility of read counts between PhoRC biological replicates. For all three samples: mesoderm-sorted Pho at either 4-6h (Pearson's correlation coefficient of 0.90; left-panel), or 6-8h of embryonic development (Pearson's correlation coefficient of 0.67; middle-panel), and whole-embryo dSfmbt ChIP (Pearson's correlation coefficient of 0.92; right-panel).

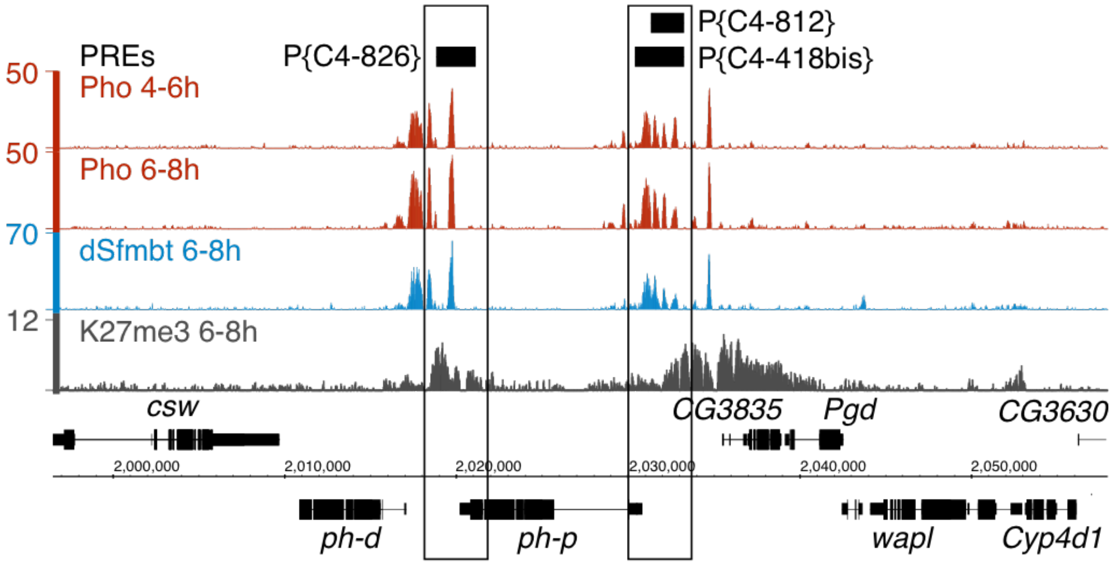

Figure 13. PhoRC occupancy on known Polycomb Response Elements (PREs) sites. Apart from the Pho and dSfmbt ChIP samples, a repressive histone mark H3K27me3 deposited by the methyltransferase in the PRC2 group is shown. Read counts have been RPGC-normalized and input (ChIP) or H3-subtracted (epigenetic marks).

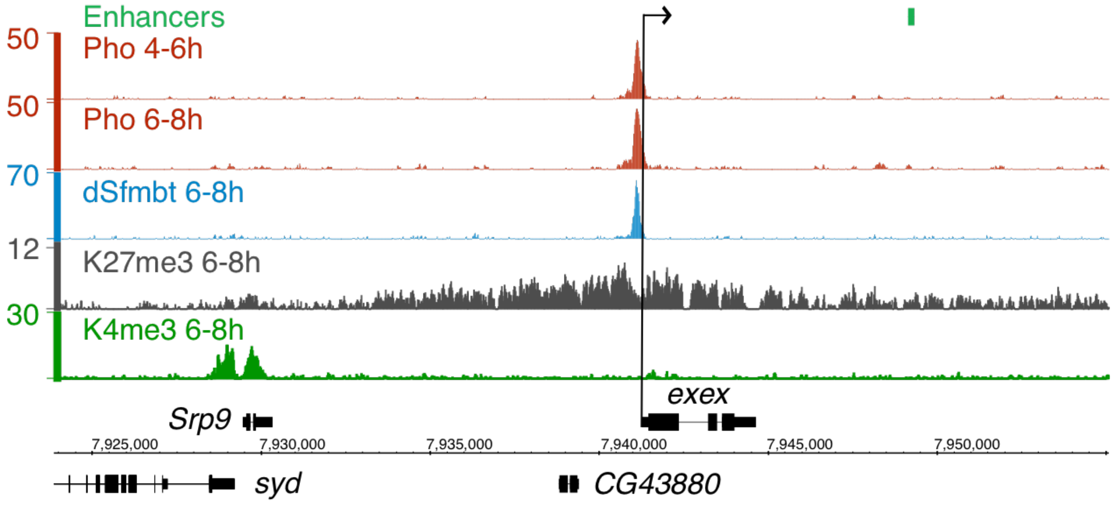

Figure 14. PhoRC occupancy on the ectodermally expressed exex promoter (Liu, J. et al. 2008). Histone mark H3K4me3 representing the active promoters is shown in dark green colour.

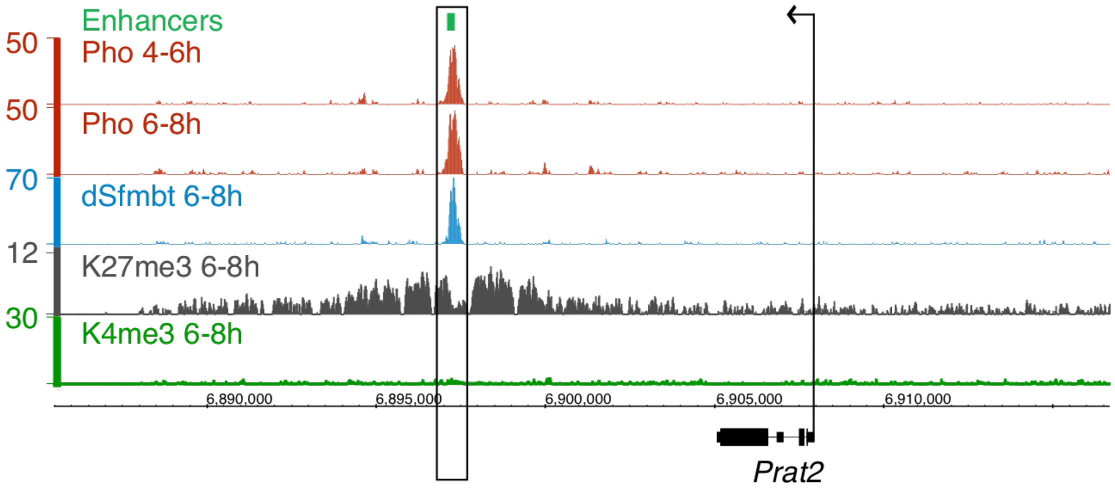

Figure 15. Newly discovered PhoRC occupancy on developmental enhancer. Enhancer is located in the vicinity of Prat2 gene, which is expressed in yolk nuclei at corresponding embryonic stage (11-12; 6-8h; Malmanche, N. et al. 2004).

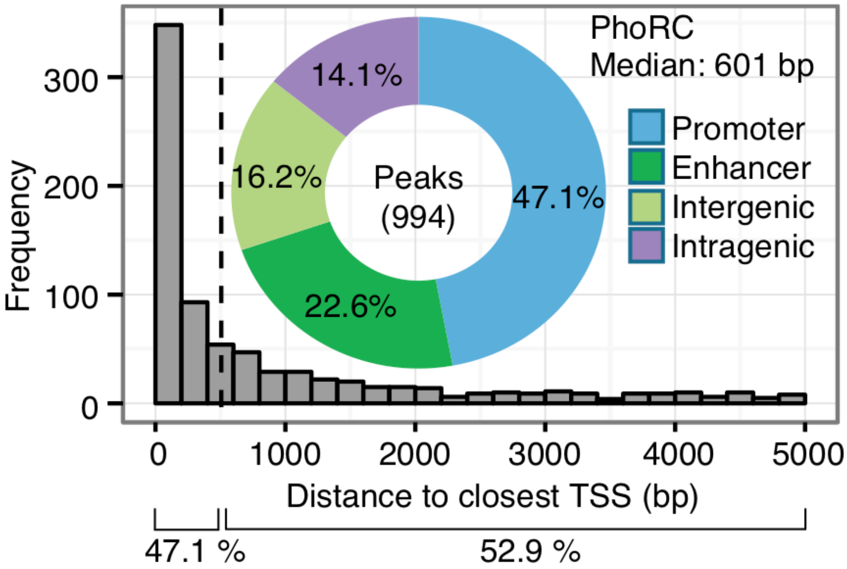

Figure 16. Distribution of distances from PhoRC peak summits to the closest transcription start site (TSS). Doughnut plot shows the overlap of PhoRC sites with the four genomic elements: promoters, enhancers, intragenic and intergenic regions. Occupancy on the developmental enhancers is highly significant (4.85 Log2 odds ratio; P-value = 8.53e-60; Fisher's Exact Test) in comparison to the matched background regions (see Methods).

4.1.1.3 Characterising the binding sites of Pho and dSfmbt

Peak calling using cisGenome (see Methods) resulted in 919 and 1,068 significant peaks for Pho at 4-6h and 6-8h, respectively, and 2,461 peaks for dSfmbt at 6-8h. We defined sites occupied by the PhoRC complex as the overlapping regions bound by both Pho and dSfmbt at 6-8h (which corresponds to 994 regions). PhoRC regions are mostly located near promoters (defined as 500bp +/- around TSS, 47.1%), followed by enhancers (22.6%), intergenic and intragenic regions (16.2% and 14.1% respectively).

4.1.2 Occupancy on developmental enhancers

4.1.2.1 Occupancy and motif discovery

Given that previous studies of Polycomb function have mainly focused on promoter regions and PREs (Figure 14; Schuettengruber, B. et al. 2009, Oktaba, K. et al. 2008, Kwong, C. et al. 2008, Schwartz, Y. B. et al. 2006), I found the high percentage of overlap between PcG peaks and developmental enhancers (Figure 15; 22.6%) intriguing. To evaluate the significance of Pho localization on the defined set of developmental enhancers, I constructed a background set of regions that matched

Table 1. Descriptive statistics for PhoRC biological replicates. Description of the encoding, read counts, filtering process, shift estimates, correaltion coefficients and called peaks for the individual biological replicates of the three ChIP samples (Pho 4-6h and 6-8h, dSfmbt at 6-8h) used in the study.

Sample	Sequenced Reads	Uniquely Aligned	Multi-mapped Reads	Failed Reads	Duplicate Percentage	Post-duplication Removal Read Count	Shift Estimates (bp)	Pearson's correlation coefficient	Peaks
dSfmbt 6-8h R1	13,476,376	9,741,380	3,077,429	657,567	78.72%	2,072,604	80	0.92	2,461
dSfmbt 6-8h R2	12,300,670	9,672,203	2,027,205	601,262	77.15%	2,210,313	75		
Pho 4-6h R1	34,301,966	24,285,572	5,838,634	4,177,760	76.93%	5,603,596	85	0.90	1,068
Pho 4-6h R2	9,753,965	6,879,538	1,339,694	1,534,733	30.39%	4,788,927	80		
Pho 6-8h R1	28,132,629	20,344,326	4,016,300	3,772,003	84.02%	3,250,128	90	0.67	919
Pho 6-8h R2	45,524,791	33,061,875	6,582,198	5,880,718	77.24%	7,524,973	85		

genomic properties such as GC content, mappability, TSS proximity and chromatin accessibility of the observed set (see Methods). When I compared the frequency of overlap of PhoRC sites on enhancers (Figure 16; 225 out of 994, 22.6%) to the one in the most penalising selection of background regions (10 out of 994, 1%), I find an astonishing enrichment of 4.85 Log2 odds ratio, resulting in a P-value of 8.53e-60 (Fisher's Exact Test). This result indicates that PhoRC occupancy on enhancers is not due to chance (for example, simply due to the fact that enhancer regions cover portion of the genome), but rather a targeted occupancy, and is likely an underestimate, since a number of regions without overlap to known enhancers (labeled as 'intergenic') contain high levels of the epigenetic mark H3K4me1, and therefore might represent uncharacterised regulatory regions.

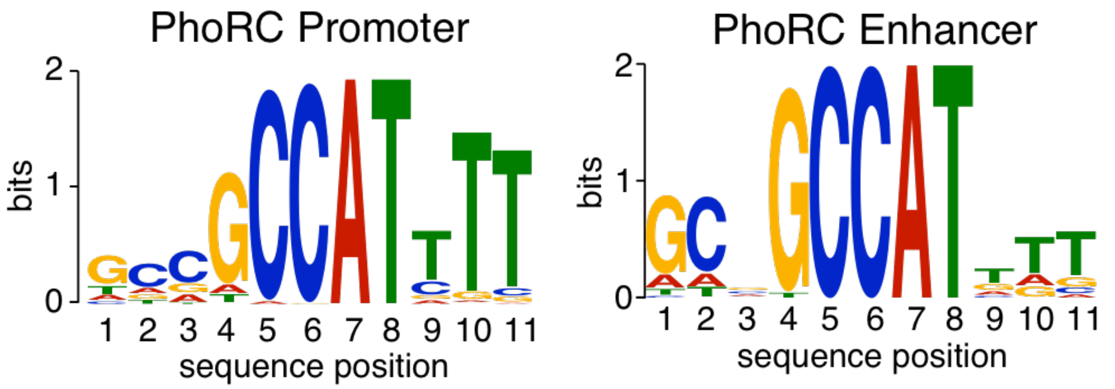

Figure 17. *De novo* motif discovery underlying PhoRC peaks. Performed on PhoRC promoter (left-panel), or enhancer (right-panel) peaks using MEME (see Methods). Motif content is matching the known Pho binding sequence (GCCAT), and is consistent on both genomic features.

4.1.2.2 Discovery of *de novo* binding motifs

Pho directly binds to DNA through its zinc finger protein domain, and has a well-described binding motif (GCCAT; Oktaba, K. et al. 2008). Using the tool MEME (see Methods), I performed *de novo* motif analysis in order to compare the motif composition between promoter and enhancer bound PhoRC sites. Although there is slight variation between the two motifs, most of the core motif is exactly matched (Figure 17), suggesting that the PhoRC complex is directly recruited to enhancers by directly binding to DNA, in a similar manner as previously described on promoters. In addition, both dSfmbt and Pho have the highest levels of RPGC-normalized read counts (occupancy) on enhancer regions, indicating that the newly discovered overlap between PhoRC binding sites and developmental enhancers is not due to low level spurious occupancy. This is further demonstrated by the increased occupancy when both proteins are in complex, rather than binding alone (Figure 18). Taken together, I found strong evidence that a significant number of PhoRC sites are located on developmental enhancers, contain a DNA-binding motif for the Pho zinc finger, and also show the highest levels of occupancy compared to any other genomic feature, including promoters (Figure 19). This raises the question of what is the function the PhoRC complex on enhancer elements?

4.1.3 Functionality of repressed enhancers

4.1.3.1 Co-occupancy of PhoRC with other PcG proteins

To examine if PcG enhancer occupancy has a functional role at enhancers, I compared their properties to promoter elements. I included other datasets such as PRC1 binding sites or histone modifications that reveals the activity state of both enhancers and target promoters in my analysis. I used publicly available ChIP datasets from whole-embryos at matching developmental stages describing the binding of the PRC1 proteins Polycomb (Pc) and Polyhomeotic (Ph), alongside two other potential alternative recruiters of Polycomb repressive complexes than Pho and Sfmbt, which are Dorsal switch protein 1 (Dsp1) and GAGA factor (Gaf) (reviewed in Müller and Kassis, 2006). When I compared the frequency of overlap of PhoRC regions to each one of TFs mentioned above, I found that the two PRC1 proteins - Pc and Ph - are highly enriched on enhancers over repressed promoters (Figure 21; Log2 odds ratios of 1.39 and 0.79, respectively), and these enrichments are statistically significant (P-value of 3.24e-06 and 4.43e-03; Fisher's Exact Test). In total, 138 out of 225 PhoRC-bound enhancers are also bound by Ph, while Pc impressively overlaps with 78% of PhoRC-bound enhancers (175). 136 developmental enhancers (60.4%) contain all

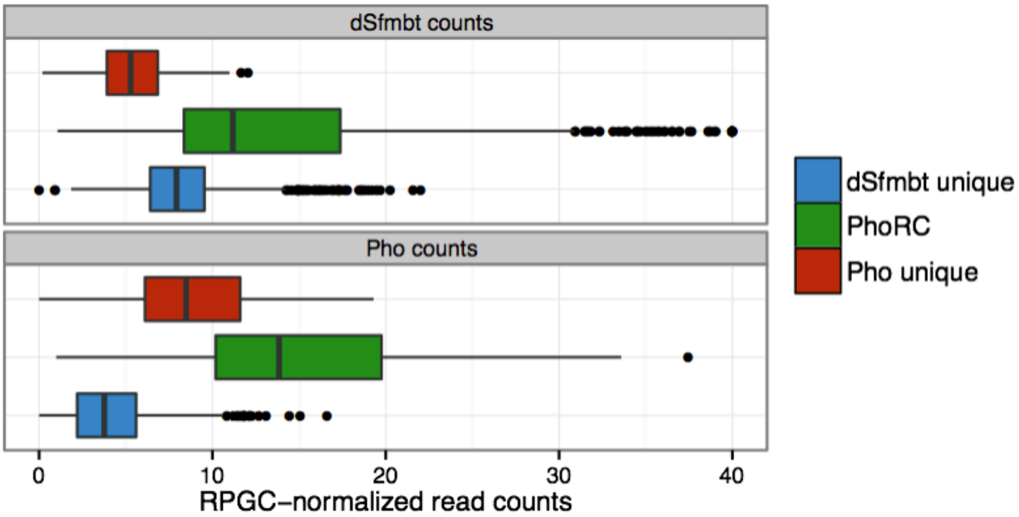

Figure 18. Distribution of dSfmbt and Pho read counts. Both contains samples from 6-8h of development summarised across the three types of peaks: unique Pho and dSfmbt loci (red and blue color respectively), or overlapping regions (PhoRC; in green).

four Polycomb-related proteins: Ph, Pc, Pho and dSfmbt. This result suggests that the PhoRC complex on enhancers is more likely to be co-occupied by functional repressors from the PRC1 complex compared to the classically described promoter regions.

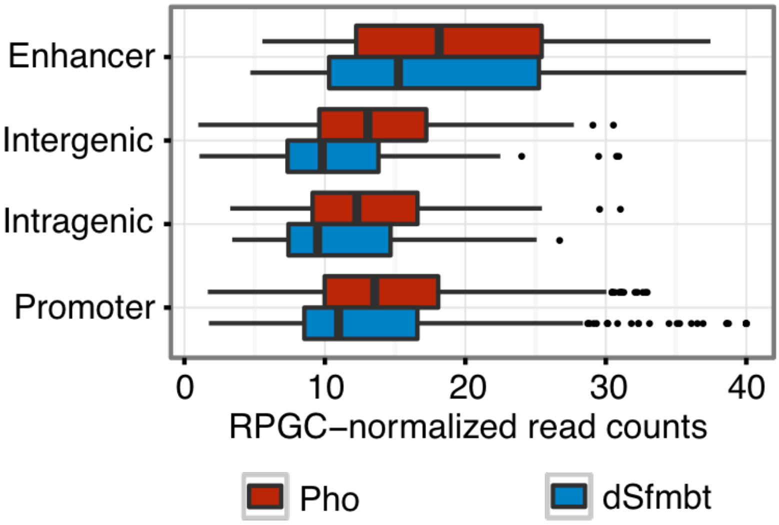

Figure 19. Read counts for PhoRC on different genomic features. Distribution of normalised read counts of Pho (in red) and dSfmbt (blue; both at 6-8h developmental stage) for PhoRC peaks that overlap one of the four genomic features: enhancers, intergenic regions, intragenic loci and promoters.

4.1.3.2 Spreading of the repressive H3K27me3 histone mark

Post-translational modification of lysine 27 of histone H3 (H3K27) by addition of three methyl groups is deposited by the PRC2 complex and this chromatin modification is essential for PcG mediated repression (Pengelly et al., 2013). Methylation of H3K27 is effectuated by a sole methyltransferase called Enhancer of Zeste (E(z)) within PRC2 – another major PcG protein complex. In contrast to the sharp localization of PRC2 at discrete genomic sites, H3K27me3 often spreads across wide regions, sometimes even up to 100kb from the origin of PRC2 recruitment. I have used a ChIP-Seq dataset of H3K27me3, previously generated in our lab from mesodermal cells at 6-8h of embryonic development (Bonn, S. et al. 2012) to describe the epigenetic state at PhoRC peaks on enhancers and promoters. Using a heatmap representation of summarized H3K27me3 signals around PhoRC peaks, it is clearly

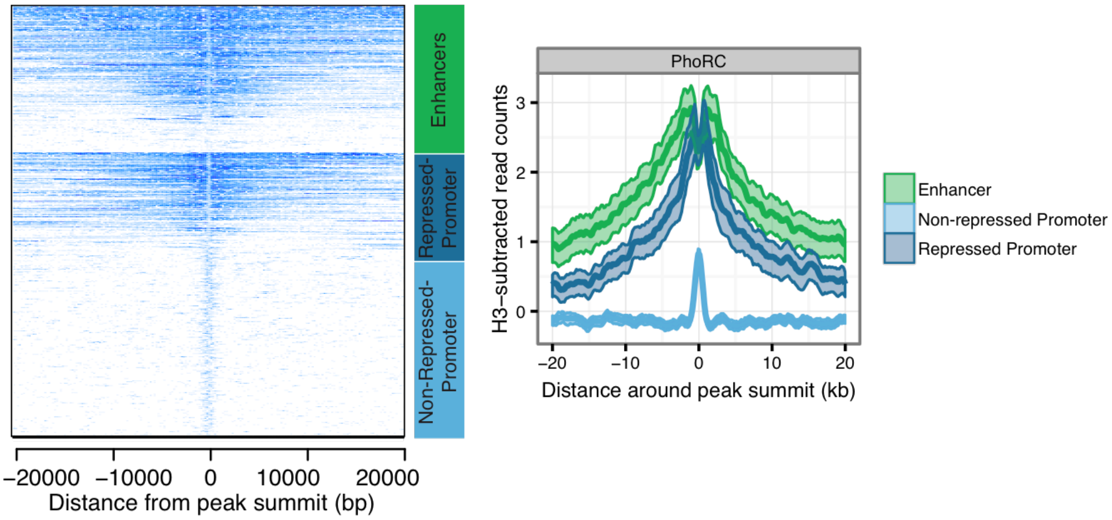

Figure 20. Spread of the repressive epigenetic histone mark H3K27me3 from developmental enhancers. This mark is deposited through PRC2 methyltransferase activity. H3-subtracted H3K27me3 signal (each row; left-panel) extends for more than 20kb from PhoRC peak summits, split between the overlap with enhancers and promoters. Additionally, promoter regions have been grouped into repressive-, and non-repressive, depending on the amount and shape of H3K27me3. Summary of the heatmap is shown on the right, where high quantitative levels of repressive mark are visible on enhancers, as well as the strictly localised extension on non-repressed promoters. Shadings indicate 95% confidence interval from bootstrap estimation.

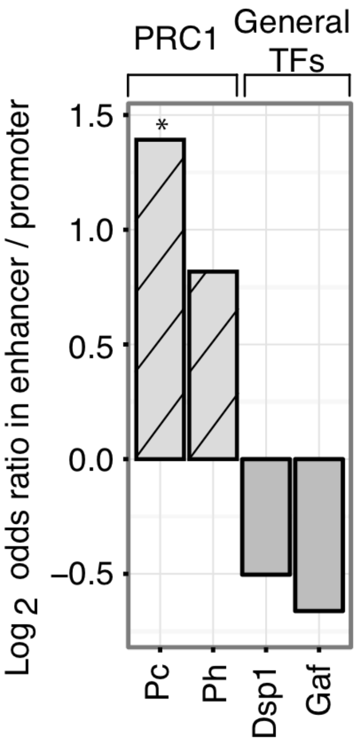

Figure 21. Enrichment of PRC1 occupancy in developmental enhancers over repressed promoters. Shown is log odds ratio from the comparison of proportions of occupancy of PRC1 components Polycomb (Pc) and Polyhomeotic (Ph) on developmental enhancers over promoter-bound PhoRC peaks, alongside the general TFs that were associated with the PcG co-recruitment. Star indicates the significance at 5% P-value (Fisher’s Exact Test).

visible that the majority of developmental enhancers overlap with wide H3K27me3 domains, while promoters exist in two distinct states, with and without the presence of H3K27me3 (Figure 20). On 46.6% of promoters, a widespread signal of H3K27me3 is present (218 out of 468), similar to what is observed at enhancers, while on the remaining promoters, there is a very localised H3K27me3 signal, which I refer to as repressed-, and non-repressed- promoters, respectively. A broad H3K27me3 distribution covers large portions of the genome, including known targets such as the *Antennapedia* (*Antp*) and *Ultrabithorax* (*Ubx*) loci, which are Homeobox (Hox) genes that are directly involved in body plan patterning (Sparmann, A. & Van Lohuizen 2006). On enhancers and repressed promoters there is also a lower H3K27me3 signal intensity directly at the site of PhoRC occupancy, which reflects the depletion of nucleosomes at PREs (due to direct binding of Pho to DNA), as observed previously (Schwartz, Y. B. et al. 2006; Mohd-Sarip, A. et al. 2005). However, this in contrast is not observed at nonrepressed-promoters, suggesting that the nucleosome might still be present at the site of the PhoRC peak summit at these promoters. Smoothed averages of the H3K27me3 signal even more clearly demonstrate the similarity between enhancer and repressed-promoter H3K27me3 profiles, suggesting that a similar Polycomb repressive mechanism might occur at developmental *cis*-regulatory modules. Apart from marking repressed Polycomb target genes, H3K27me3 is also a characteristic signature of repressed enhancer activity and antagonizes acetylation of the same lysine (H3K27ac) – a mark associated with enhancer activation. I have used information on the binding of 5 mesodermal TFs (Twist (Twi), Myocyte-enhancing factor 2 (Mef2) and Tinman (Tin), Biniou (Bin) and Bagpipe (Bap)) to categorize developmental enhancers into two classes: PhoRC bound, and PhoRC unbound (with at least 2 meso-TFs; Figure 22). Reflecting the genome-wide pattern, PhoRC occupancy on developmental enhancers themselves correlates with high levels of the repressive H3K27me3 mark (Figure 23), while the absence of PhoRC accompanied by the presence of meso-TF binding shows the opposite enrichment of the H3K27ac active enhancer mark. This result suggests that PhoRC on developmental enhancers recruits PRC2 components that deposit H3K27me3, thus becoming repressed. On the other hand, lack of PhoRC alongside the binding of mesodermal TFs leads to the enhancer activation.

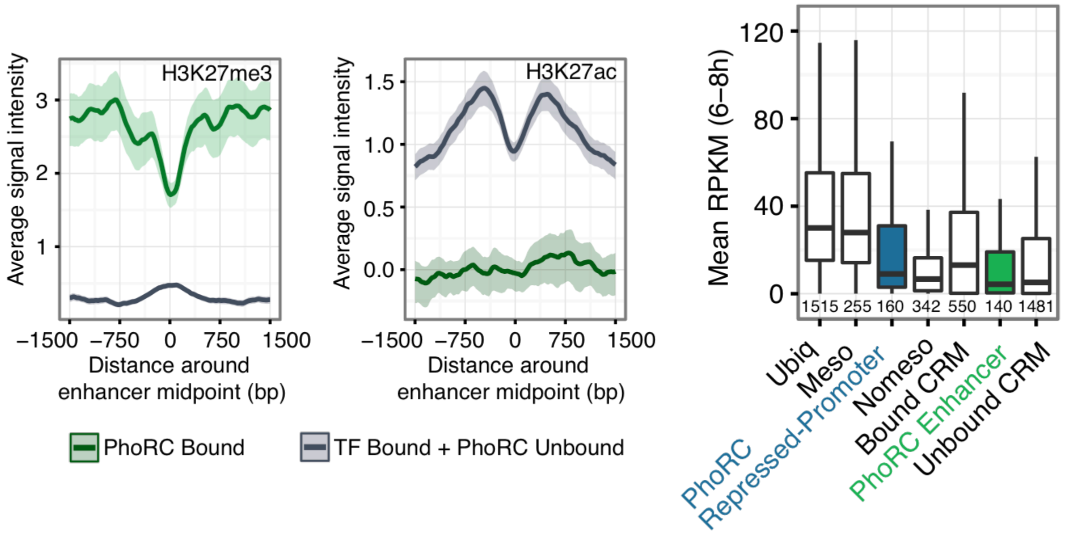

Figure 22. Activity states on the PhoRC-bound developmental enhancers. Represented by the histone state, including representative marks of activity (H3K27ac), and repression (H3K27me3). Shadings indicate the 95% confidence interval as estimated by the bootstrap. In green PhoRC peaks are shown, while in grey the rest of the developmental enhancers from the list that were bound by at least two other transcription factors. On right, RPKM values from mesoderm-sorted RNA-Seq experiment as compared between reference groups, and neighbouring genes of PhoRC bound loci.

4.1.3.3 Activity of target genes

Another way to assess a potential function of PhoRC enhancer binding, is to assess the transcriptional state of the neighboring gene (Figure 22). For this, I used mesoderm-specific RNA-Seq data (Gaertner, B. et al. 2012) from the same developmental stages as the ChIP experiments. As a reference for comparison with our list of genes, I categorized the genes into different classes, based on their spatial patterns of expression, using in-situ hybridization data: 'ubiquitous' (expressed through the embryo), 'meso' (gene expression in mesoderm and potentially other tissues, but not ubiquitous), and 'non-meso' (expression lacking mesodermal terms, but not ubiquitous), along with the two classes of enhancer occupancy that have been shown to correspond to the activity: 'Bound CRM' having two or more associated meso-TFs, and 'Unbound CRM' having no meso-TF occupancy at the 6-8h of embryonic development.

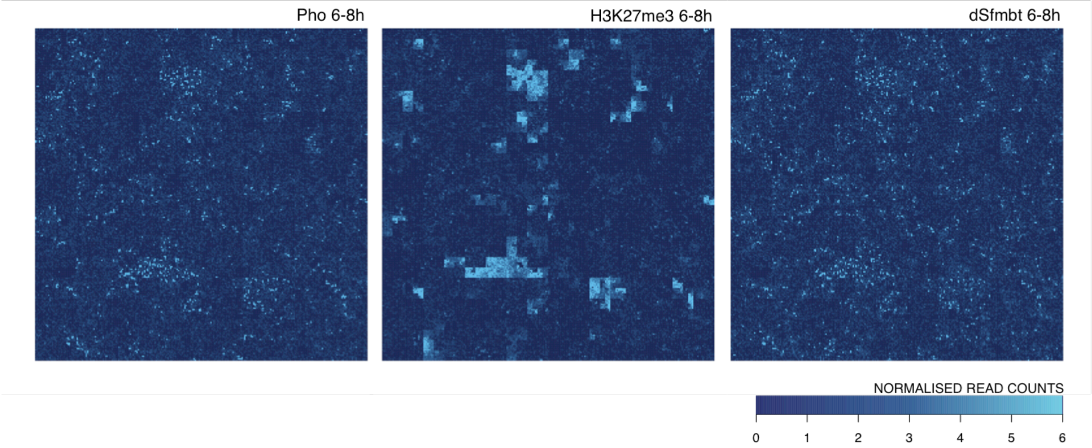

Figure 23. Hilbert curve representing PhoRC occupancy on chr3L. Map of binding for three different factors: Pho, H3K27me3 and dSfmbt on chromosomal arm 3L. Linear coverage is represented using 8-fold Hilbert curves, and the intensity of the color indicates the varying levels of occupancy.

Genes closest to the PhoRC peaks on repressed promoters have significantly lower levels of expression (RPKM values) compared to active genes, with the median equivalent to genes expressed outside the sorted tissue. More interestingly, genes that are in the vicinity of PhoRC-occupied developmental enhancers exhibit the same pattern, having significantly lower levels of transcription compared to genes in the neighbourhood of meso-TF bound CRMs. Overall, these results show that PhoRC occupancy on development enhancers leads to recruitment of PRC1 and PRC2 proteins, which in turn deposits high levels of the characteristic Polycomb repressive histone mark H3K27me3, spreading more than 10kb from the initial recruitment site occupied by Pho and dSfmbt, and likely results in low levels of transcription of neighboring genes.

4.1.4 Dynamics of Pleiohomeotic occupancy

4.1.4.1 Change in occupancy during developmental progression

Polycomb functionality was originally associated with the long-term maintenance of gene repression in fruit flies (Lewis, E. B. A 1978), however comparisons were usually made between very large time spans (such as embryos to imaginal discs, or tissue culture cell lines) or only on specific regions that were previously linked to Polycomb binding. Due to our unique experimental design, I had the opportunity to

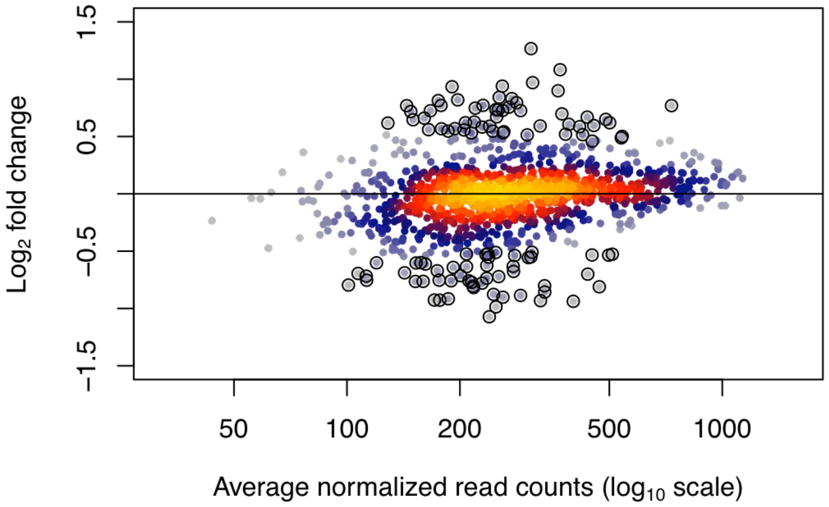

Figure 24. Differential Pho occupancy between 4-6h and 6-8h of embryonic development. Density of the points is indicated by the warmer colors. Significant differences (with less than 10% false-discovery rate and at least 1 fold change) are marked in black circles.

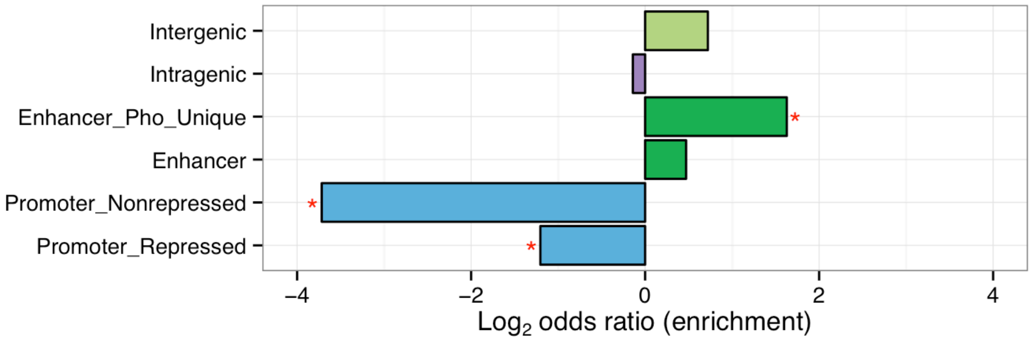

Figure 25. Enrichment of genomic features overlapping differential PhoRC occupancy sites. Red stars indicate the significant (less than 5% P-value; Fisher's Exact Test) increase or decrease in proportion of differential peaks on a specific feature compared to the genome-wide distribution.

quantify the dynamic occupancy of Pho genome-wide (including promoters and developmental enhancers) at short developmental progression (4-6h to 6-8h, corresponding to stages 7-9 to 10-12). To perform this analysis, I used DESeq2, a tool based on the negative binomial distribution (see Methods). Shifted Pho read counts (which then represent the true loci of protein occupancy) have been counted in +/-2kb regions around 1,198 PhoRC and Pho-unique peak summits. Using a default set of DESeq2 parameters (library size factor normalization, parametric estimate of dispersion and binomial wald test with independent filtering), I found 117 regions to have a significantly different occupancy between developmental stages at a 10% false-discovery rate (9.56% of total peaks; Figure 24) - reflecting the remarkable stability of Polycomb protein binding even between dramatic developmental transitions (before and after mesodermal specification). Interestingly, I found that promoter-overlapping peaks, and especially the ones without the H3K27me3 mark, tend to be quite static, while dynamic Pho occupancy on enhancers is higher than expected (Figure 25; compared to the overall proportion of enhancer peaks within the whole set), but only when Pho is not co-occupied with dSfmbt. Unique Pho occupancy has a highest average fold change between conditions (median value of 0.3), which is significantly higher than the closest category of intergenic peaks (P-value = 0.018; Two-sided Mann-Whitney U Test). Overall, these results suggest that PhoRC occupancy to enhancers, as well as to repressed-promoter regions, is very stable and static during developmental progression, a result that fits PhoRC's known role in maintaining gene repression over longer developmental spans. In contrast, enhancer occupancy by PhoRC seems to more responsive and following the transitions in developmental phenotype.

4.1.4.2 Spatiotemporal co-occupancy with the master regulator of mesoderm development Twist

Developmental enhancers are commonly bound by a range of TFs that either prime the site by displacing the nucleosomes (pioneer TFs), recruit other factors, or directly regulate target gene expression (Spitz, F. & Furlong, E.E.M. 2012). I found that the general TFs that were previously suggested to be additional recruiters of Polycomb complex (Dsp1 and Gaf) are depleted on enhancers. I therefore wondered if there are any other potential TFs on enhancers that significantly co-localize with Polycomb

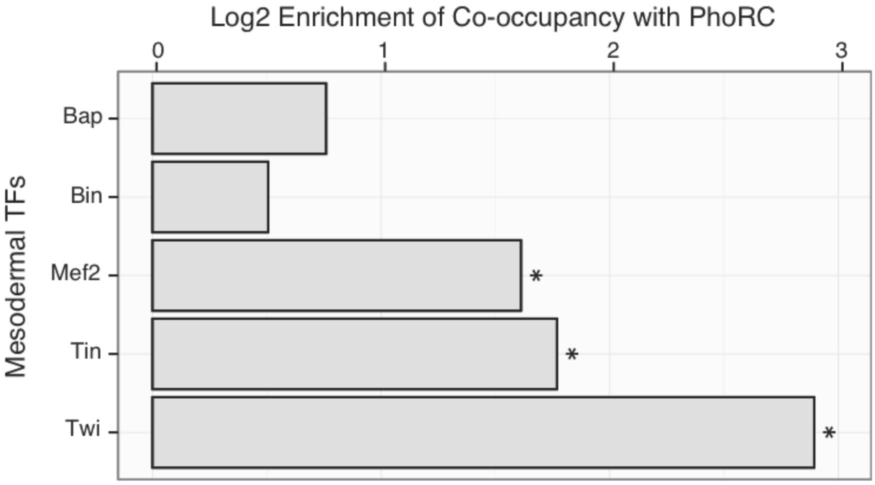

Figure 26. Co-occupancy of meso factors with PhoRC. Enrichment of occupancy of mesodermal transcription factors Bap, Bin, Mef2, Tin and Twi with the PhoRC complex on developmental enhancers. Odds ratio represents the increased proportion compared to the genome-wide levels of TF8008 occupancy. Stars indicate the significant enrichment with false-discovery rate lower than 10%.

proteins. The majority of PhoRC occupied developmental enhancers (180 out of 225) overlap with previously identified TF-defined CRMs (TF8008), which are based upon the occupancy of five mesodermal proteins, ranging from the master regulator of mesodermal development Twist to the ones controlling the specification of particular mesodermal sub-tissues, such as Tinman for cardiac tissues and Bagpipe for visceral muscle (Zinzen, R. P. et al. 2009). Since the binding profiles of these mesodermal factors where performed at equivalent developmental stages, I analysed the overlap of Pho with each mesodermal TFs. Likelihood analysis of co-occupancy between Meso-TFs and PhoRC sites revealed a significant enrichment for all three major mesodermal regulators: Mef2, Tin and Twi (Figure 26). Twist is a bHLH protein that is functionally equivalent to MyoD in vertebrates, and shows the highest enrichment with the PhoRC complex (Log2 odds ratio of 2.89, P-value = 6.13e-11; Fisher's Exact Test). Although no particular motif grammar between Twi and PhoRC was found, the strong co-occupancy might be a plausibly new mode for facilitating Polycomb recruitment.

I decided to further explore a possible functional interplay between the two proteins by measuring the temporal recruitment of PhoRC proteins. Based on a time-course of Twist occupancy (Zinzen, R. P. et al. 2009), I defined three Twi-dependent

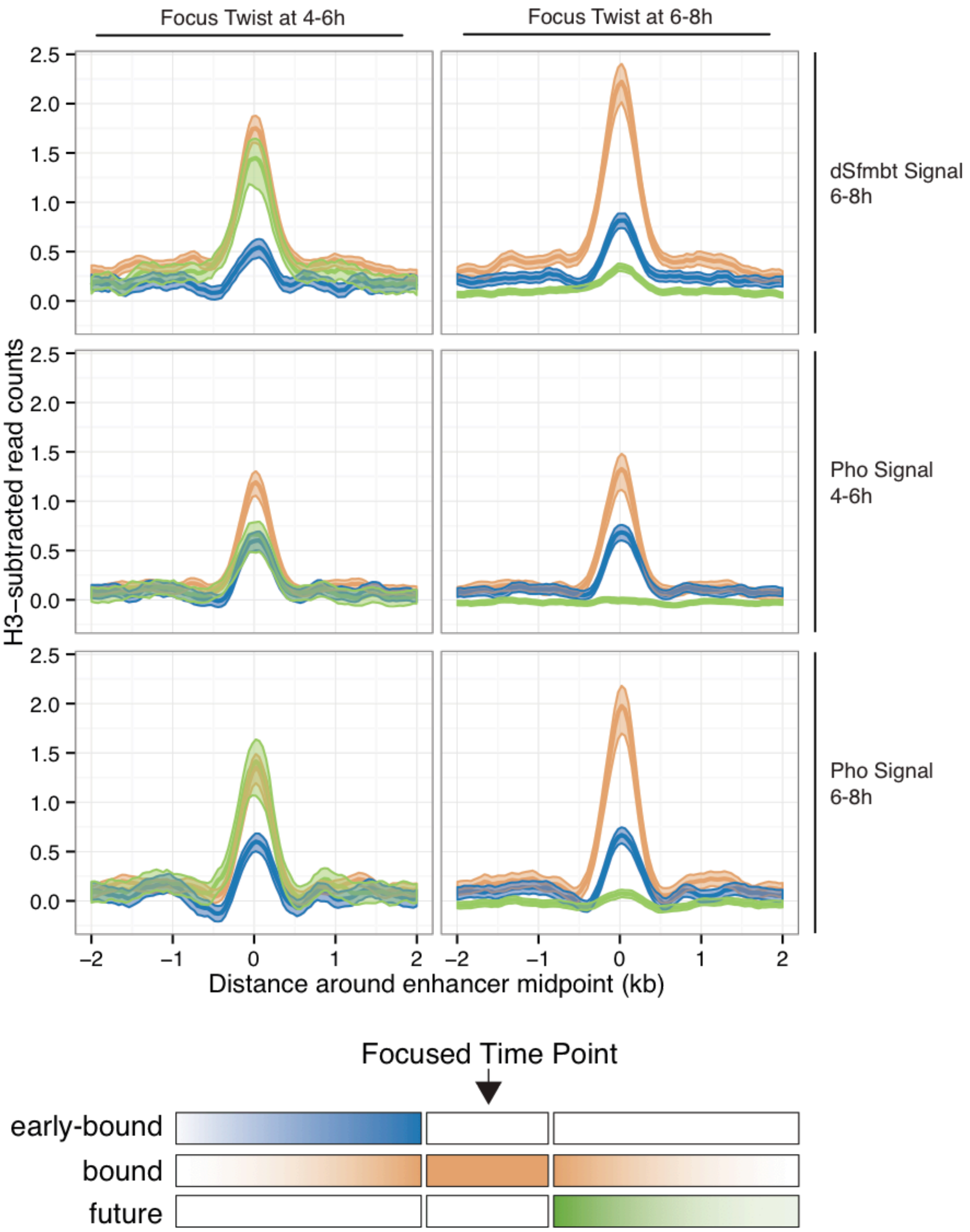

Figure 27. Temporal profiles of Pho (at 4-6h, and 6-8h of development), and dSfmbt (at 6-8h) on two Twist contexts. One where Twist binding is focused at 4-6h (left-panels), and 6-8h (right-panels). Legend below explains the three defined categories that change depending on the Twi context. Bold curves represent the 10% trimmed mean profiles, while bands indicate the 95% confidence interval around mean as estimated using bootstrap.

states: early-bound (before time point of interest, and not later), bound (at the time point of interest), and future (only after the specified time point), with the bound focus point being at either 4-6h or 6-8h of embryonic development (Figure 27). For example, at 4-6h I define 'early-bound' category as all the enhancers that have Twi bound only at 2-4h, but not 4-6h and 6-8h. In all cases Pho, either from the 4-6h or 6-8h dataset, as well as dSfmbt at 6-8h show clear correlations and higher quantitative signals when Twi is bound. Interestingly, in the majority of cases there is a lower, but significant amount of Pho and dSfmbt occupancy at enhancers where Twist has just become unbound, indicating the possibility of TF perdurance after the initial co-localisation. Most intriguing are the instances of 'future' Twi binding, where at 4-6h, strong enrichment of both Pho and dSfmbt is clearly visible before the onset of the next developmental time point, indicating that Twi might act as a stabiliser of repressive occupancy, rather than a pioneering factor in the case of Polycomb recruitment. Before, Polycomb proteins were associated with long-term repression (reviewed in Lande-Diner, L. & Cedar, H. 2005). Our results however suggest that the binding of PhoRC is much more dynamic on developmental enhancers, priming the loci that are bound by mesodermal factor Twist at the later stage of development.

4.1.5 Modes of Polycomb mediated repression in mammals

4.1.5.1 Enrichment of the vertebrate homolog of Pleiohomeotic (Yin Yang 1) on human enhancers

Discovery of PhoRC occupancy on developmental enhancers in *Drosophila melanogaster* might have a significant impact on our understanding of the biology of repressive Polycomb regulation during development. All Polycomb group proteins have mammalian homologues (detailed in the introduction chapter), with some, such as the recruiter Pho showing 96% protein identity with the human protein Yin Yang 1 (YY1; Brown, J. L. et al. 2003). Using publicly available data from the ENCODE consortium, I tested the likelihood of occupancy of YY1 on putative human enhancers that were identified based on the DNaseI hypersensitivity signal (ENCODE Project Consortium 2012; Ho, J. W. K. et al. 2014). Comparing the proportion of overlap between observed YY1 peaks, and the matched background set (see Methods) revealed two findings: 1. YY1 ChIP-Seq in all cell lines examined (7, with the exception of K562 and SK-N-SH cells), show a significant enrichment at DNase I hypersensitive sites (median Log2 odds ratio of 0.54; Figure 28); 2. YY1 seems binds at many regions in a cell type-specific manner. For example, ChIP of YY1 in human embryonic stem cells (H1-hESC) shows the highest enrichment over background (H1; 1.56 Log2 odds ratio; P-value < 1e-40; Fisher's Exact Test), similar to the results in the lymphoblastoid cell line GM12878 (2.31 Log2 odds ratio; P-value < 1e-40; Fisher's Exact Test) and in the erythroleukemic cell line K562 (0.83 Log2 odds ratio; P-value < 1.3e-22; Fisher's Exact Test). Overall, the results showed that the YY1 DNA-binding homologue of the Polycomb recruiter Pho shows significantly enriched cell-specific occupancy on human enhancers, similar to our finding in flies, indicating that this observation is evolutionary conserved between insects and mammals.

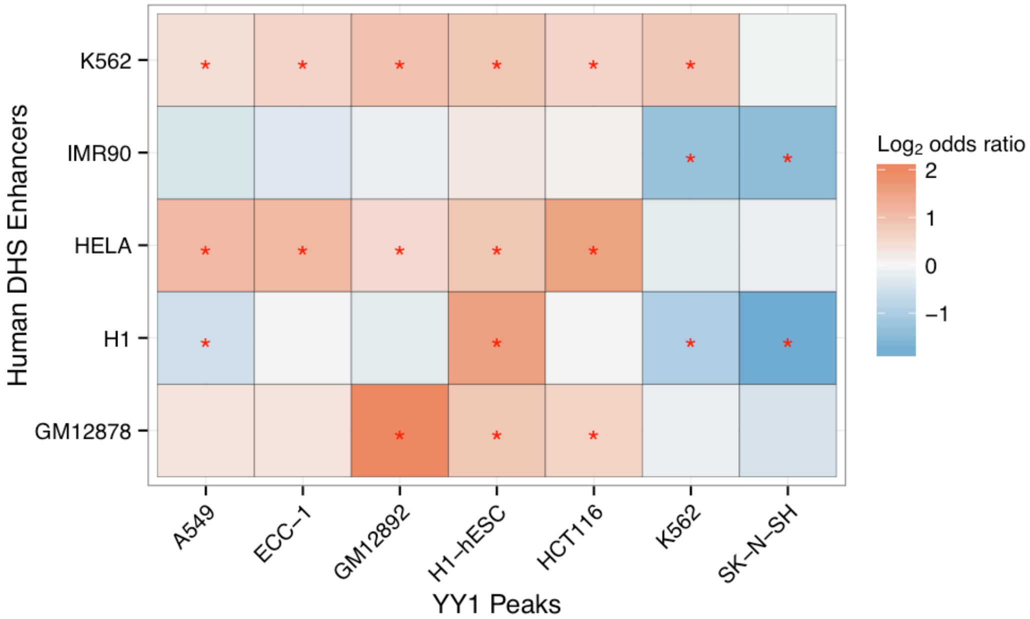

Figure 28. Enrichment of YY1 occupancy in different cell types. YY1 from various cell types (x-axis), was tested against the matched background (see Methods) on human enhancers defined by the DNaseI-hypersenstivity (Y-axis). Gradient of colours represents the enrichment (warm), or depletion (cold) of the YY1 compared to expectation. Stars represent the significant observations (<1e-05 adjusted P-value; Fisher's Exact Test, adjusted using Bonferroni method).

4.1.5.2 Representation of epigenetic marks on human developmental enhancers reveals genomic signatures reminiscent of Polycomb repression in flies

Following the results showing an increased and significant occupancy of YY1 on human enhancers, I used ENCODE data on histone modifications that represent enhancer sites (H3K4me1), repressed transcription (H3K27me3), and potentially

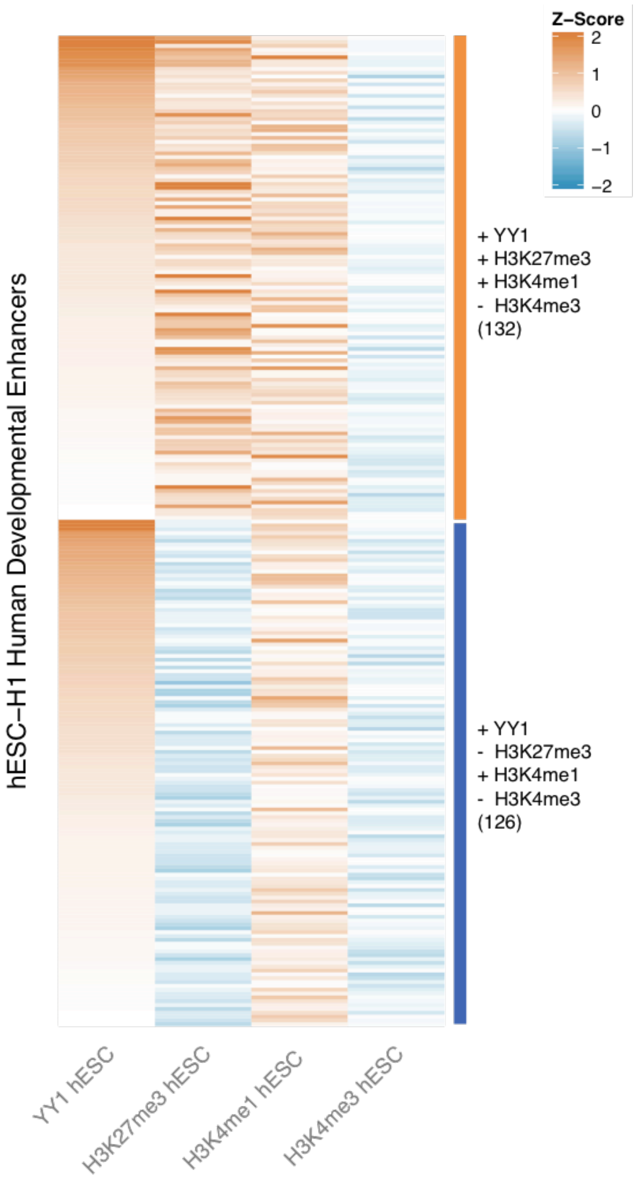

Figure 29. Epigenetic signatures on YY1-bound developmental enhancers in humans. Ranked occupancy of vertebrate homologue of Pho, Yin Yang 1 (YY1) in two clusters defined by the histone marks representing enhancer locations (H3K4me1), repression (H3K27me3) and potentially unannotated promoters (H3K4me3) on human developmental enhancers (from Rada-Iglesias, A. et al. 2011). Clusters are divided in groups with (orange), or without (blue) H3K27me3 signal, reflecting the dual functionality of YY1.

unannotated promoters (H3K4me3) to functionality characterise YY1 binding on a different set of enhancers from those described in the previous paragraph. These were defined based on the occupancy of P300, BRG1 and H3K4me1 during human ES cell differentiation (from Rada-Iglesias, A. et al. 2011). Two distinct classes of putative human developmental enhancers seem to exist: one with high H3K27me3, YY1 and H3K4me1 occupancy representing repressed enhancers (132 cases; Figure 29), and another without the H3K27me3 histone mark (126 cases). These two different enhancer states suggests that YY1-occupied enhancers may have dual regulation, reminiscent of YY1's known biological role in both transcriptional repression and activation depending on the context (Hyde-DeRuyscher, R. P. et al. 1995). Taken together, these results suggest that YY1 binds to human (putative) developmental enhancers, that bear the chromatin signatures of PcG mediated repression, similar to what we observed in flies.

4.1.6 Summary

This project determined the occupancy of the two members of the PhoRC complex in two developmental contexts: across different cell types (mesoderm specific using BiTS-ChIP for Pho, versus whole-embryo for dSfmbt), and at different stages of development (at 4-6h and 4-6h of *Drosophila* embryogenesis, marking the transition from a fairly homogeneous cell populations to their specification into different muscle primordia) to explore the effect of repressive regulation on developmental regulator genes. Our results present strong evidence of a previously unacknowledged association of the PhoRC complex with developmental enhancers. Three independent properties suggest a high significance of PhoRC occupancy on distal regulatory regions: the presence of top-ranking ChIP-Seq signals, the recovery of a DNA-binding motif for Pho, and the increased co-occupancy with PRC1 components compared to repressed promoters. PhoRC-bound developmental enhancers have a consistent and representative spread of the repressive H3K27me3 histone mark, even up to 20kb from the recruitment loci, the level of which matches the quantitative levels on repressed promoters. Genes surrounding these repressed developmental enhancers have lower levels of transcription compared to ones in the vicinity of active enhancers, suggesting that they might be repressed by the PcG system emanating from enhancer elements. Dynamic analysis of Pho occupancy across the two developmental time points revealed few differential changes, suggesting that repression is a stable process, at least between this narrow time window. Nonetheless, occupancy of the PhoRC complex seems to be strongly correlated with the temporal occupancy of the master regulator of mesodermal development, Twist. Finally, analysis of the occupancy of the human Pho homology, YY1, revealed a similar association on potential developmental enhancers, suggesting that this could be an evolutionarily conserved feature. Since mammalians seem to lack a clear genomic signatures of PREs, a next important step will be to experimentally challenge the hypothesis of a strict functional division between PREs and developmental enhancers (see Discussion) by testing the functions of known and newly defined genomic regions in standard enhancer and PRE assays *in vivo*, using *Drosophila* genetics.

4.2 Enhancer 3D interactions during embryonic development

4.2.1 Discovering significant interactions using 4C-seq

4.2.1.1 Aims of the project

Developmental enhancers regulate the expression of target genes through binding of specific transcription factors, such as the previously described Pho and dSfmbt, which recruit the repressive Polycomb group of proteins. However, although being crucial for their spatiotemporal activity, enhancers can be located far from the promoter of the gene they are regulating, often spanning large genomic distances including other genes and enhancers. Since genomic contacts between enhancers and promoters are the first step in understanding how the non-coding genome regulates metazoan development, I worked on analyzing the interaction profiles of enhancers that were previously characterized by our group (see Methods), from specific anchor sites called viewpoints, to all other loci using a chromosome conformation capture method 4C-seq. In collaboration with Yad Ghavi-Helm from our group who performed all the wet-lab experiments, including collecting the staged *Drosophila* embryos, performing the 4C protocol and preparing the sequencing libraries, and Felix Klein from Huber group who did the alignment of sequenced reads, pre-processing and modelling of statistical fit to estimate the significance of observed values, I characterised the biology of spatial interactions from more than a 100 proximal and distal enhancers in two different contexts (whole-embryo vs. mesodermal), and time-points (3-4h vs. 6-8h) in the biggest 4C-seq study up to date.

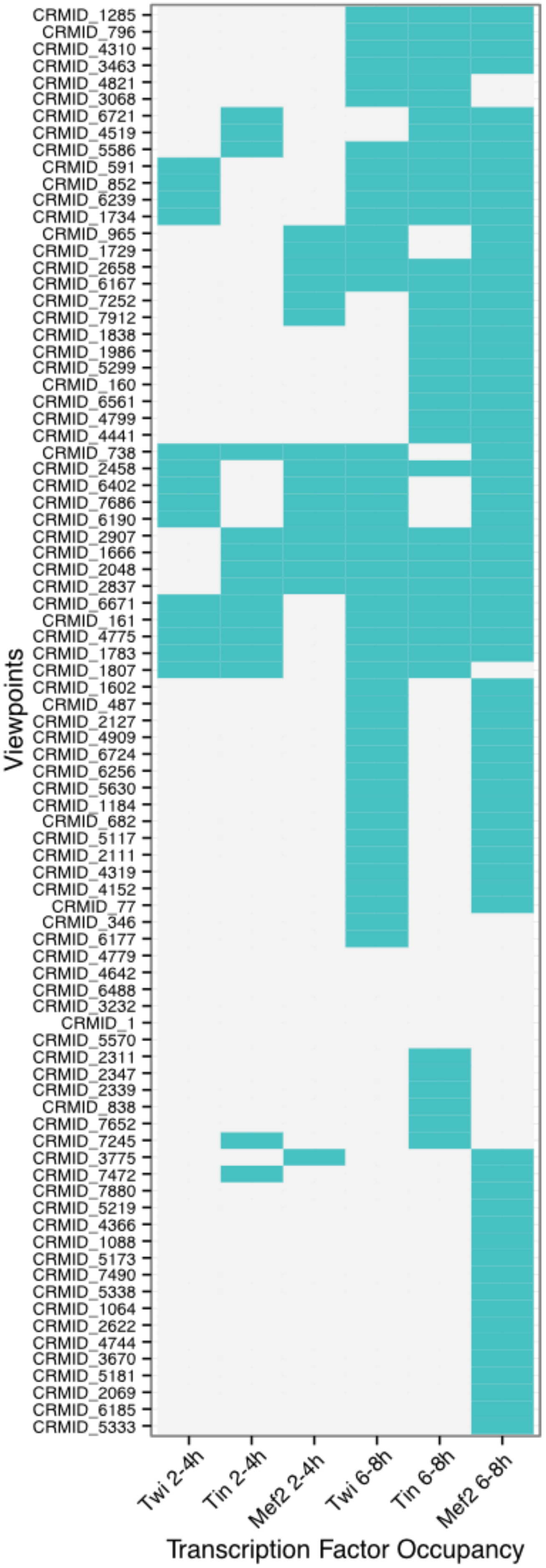

Figure 30. Occupancy of mesodermal TFs on 4C-seq viewpoints. Occupancy of three major mesodermal transcription factors - Mef2, Tin and Twi, at two development stages (2-4h, and 6-8h), on viewpoints selected for our 4C-seq experiment. Teal color represents an overlap between defined enhancer region and transcription factor peak. Matrix has been clustered using Euclidean distance and Ward's method.

4.2.1.2 Experimental design

For this experiment, 48 promoter-proximal (within 1kb of closest TSS or overlapping H3K4me3 peak) and 59 distal enhancer viewpoints were performed in four different biological contexts: 1) whole-embryo collections at developmental stages 6-7 (corresponding to 3-4h of embryogenesis at 25^0C) when cells are still multipotent; 2) whole-embryo collections when the cells are being specified into ectoderm, mesoderm and endoderm (corresponding to stages 10-11; 6-8h); 3) mesoderm-specific dataset as a result of modified 4C-seq method where fluorescently labeled nuclei driven by a promoter of a specific mesodermal marker Twist were FACS sorted at 3-4h of embryonic development (BiTS-4C-seq); 4) BiTS-4C-seq performed at 6-8h of development. This experimental design allowed us to explore the enhancers' interaction dynamics in both temporal (3-4h vs. 6-8h), and spatial (mesoderm vs. whole-embryo) contexts.

4.2.1.3 Quality control and validation of known interactions

Our design included 10 control enhancers involved in known regulatory interactions (UBX_BXD-A, Ubx_BRE, pdm2_CE8012, ap_ApME680, eya_meso_eya_1, sv7490, Con_1kb_3prime, stumps_hbr_DME, E2f, Mef2_II-BL_NR) both for the validation of data quality and for the development of quantitative statistical methods. All of the known enhancer-to-promoter interactions were recovered, including a well-characterised interaction between the apME680 enhancer, and the ap gene promoter, located 17kb away from the regulatory element. Our data showed a high degree of reproducibility, with the median Spearman's correlation coefficient of 0.93. Together with the recovery of known interactions, this result indicated that our high-resolution 4C-seq data was of both excellent specificity and sensitivity.

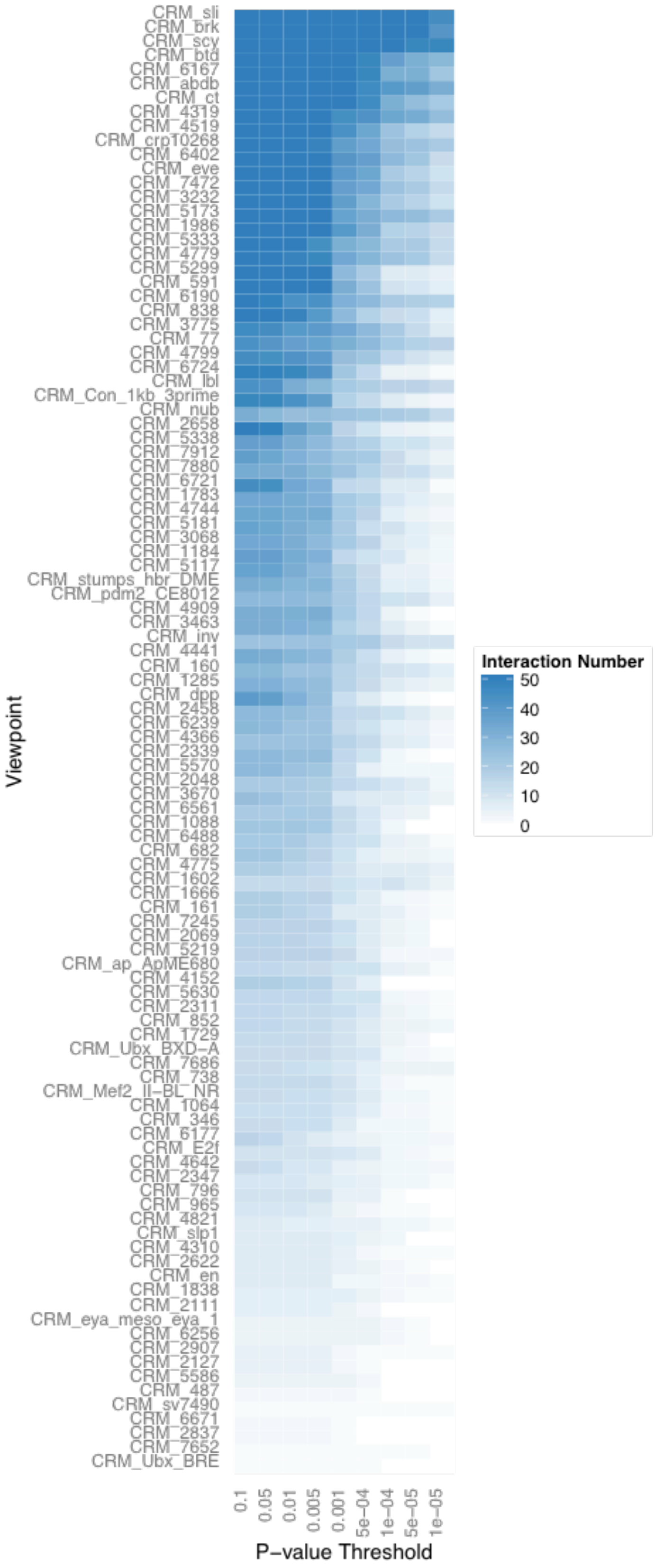

Figure 31. Heatmap of interaction frequencies for each viewpoint. Each viewpoint is sh, and range of P-value thresholds (columns) from 0.1 to 1e-05, used to define the significant interactions (see Methods). Rows have been ordered according to the median interaction number (highest to lowest).

4.2.2 Enhancer interactions during *Drosophila melanogaster* development

4.2.2.1 Viewpoints and significant interactions

Viewpoints were selected to have a wide-range of transcription factor occupancy (Figure 30) and epigenetic states, with some being bound by only a single TF from the list of mesodermal factors (e.g. CRM_2311 overlaps with only Tinman at 6-8h of development), while others are occupied by all three major mesodermal factors in both temporal contexts, such as CRM_2048. Such diversity allowed us to explore the full complexity of regulation, and it's consequence. At a P-value threshold of 0.001 (corresponding to a Z-score of 3) in both biological replicates, along with having less than 10% false-discovery rate in any of the replicates, I defined in total (over all 4 experimental conditions) 4,247 interactions (Figure 31), 1,036 of which are unique across all conditions. As we have the most extensive genomic annotations for 6-8h, I used the 1,983 significant 6-8h 4C interactions to characterize basic summary statistics for the *Drosophila* developmental enhancers interactions. On average, viewpoints interacted with 15 other regions, including 1 defined enhancer, 2 intergenic loci, 7 regions within genes, and 4 promoters. Distal enhancer viewpoints contain interactions with other developmental enhancers, while promoter-proximal enhancer viewpoints contain interactions with defined promoter regions, indicating that the spatial interactome in *D. melanogaster* is a complex topology that includes different genomic elements.

4.2.2.2 4C interactions are surprisingly distant to the viewpoints they originate from

As compared to mammalian genomes, the *Drosophila* genome is fairly compact. Of the ~16,000 genes, with ~30,000 isoforms, annotated by FlyBase v5.47, 43.05% of genes overlap at least one other gene, while 16.4% of genes are completely nested within another. The mean distance between non-overlapping, coding transcripts is 1.3kb. But despite this relative compaction, a large number of enhancer interactions were observed over substantial (linear) distances. The median interaction distance for viewpoints in our dataset was 114kb. And while the majority of known enhancers were located within 10kb of the target gene, 73% of significant 4C interactions in our dataset were found further than 50kb from their viewpoints (Figure 32). Although 4C-seq technology substantial underestimates the number of short-range contacts, the extent of long-range interactions observed here gives evidence for extensive long-

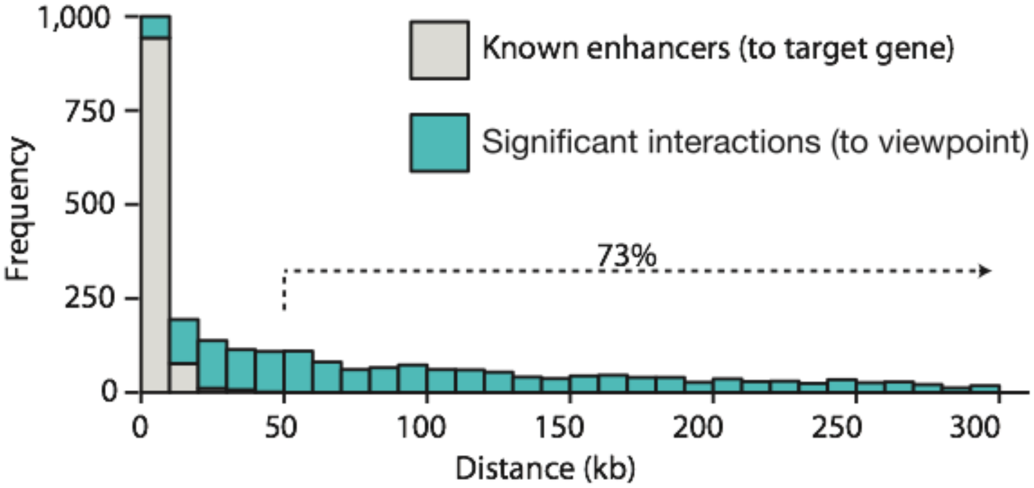

Figure 32. Distribution of viewpoint-to-interaction distances. Histogram of distances between the midpoint of significant interactions, and their corresponding viewpoints. In teal, contacts from whole-embryo 6-8h experimental conditions are represented, while grey columns contains distances of the known enhancers and their targets from previous literature sources, including CAD2 database (Bonn, S. et al. 2012), and Redfly (Gallo, S. M. et al. 2010).

range interactions, which had previously only been associated with mammalian genomes. Within our dataset, one of the longest observed interactions is more than 0.5Mb long, from the *unc-5* promoter to the promoter of *sli*, genes that share similar biological functions. The most frequent long-range contact was observed over a 235kb distance between two genes: Charybde (*chrb*) and Scylla (*scyl*), a set of paralogs that act as the inhibitors of cell growth. This described interaction can be recovered when viewpoints are placed on either *chrb*, or *scyl* loci, indicating the reciprocity and reproducibility of the local interaction space.

4.2.2.3 Interactions are contained within same topological and syntenic domains as viewpoints

When the interacting regions are overlaid with information on chromatin topology, as inferred from a published whole-embryo Hi-C experiment done at a very late stage of embryogenesis (Sexton, T. et al. 2012), it is clear that not only the Chrb-Scyl interaction, but also most of the other significant contacts within the gene desert region (long stretches of DNA with low frequency of genes) are constrained between the two paralogous genes, a range that exactly corresponds to the boundaries of the independently-defined topological domains. To test the genome-wide hypothesis that

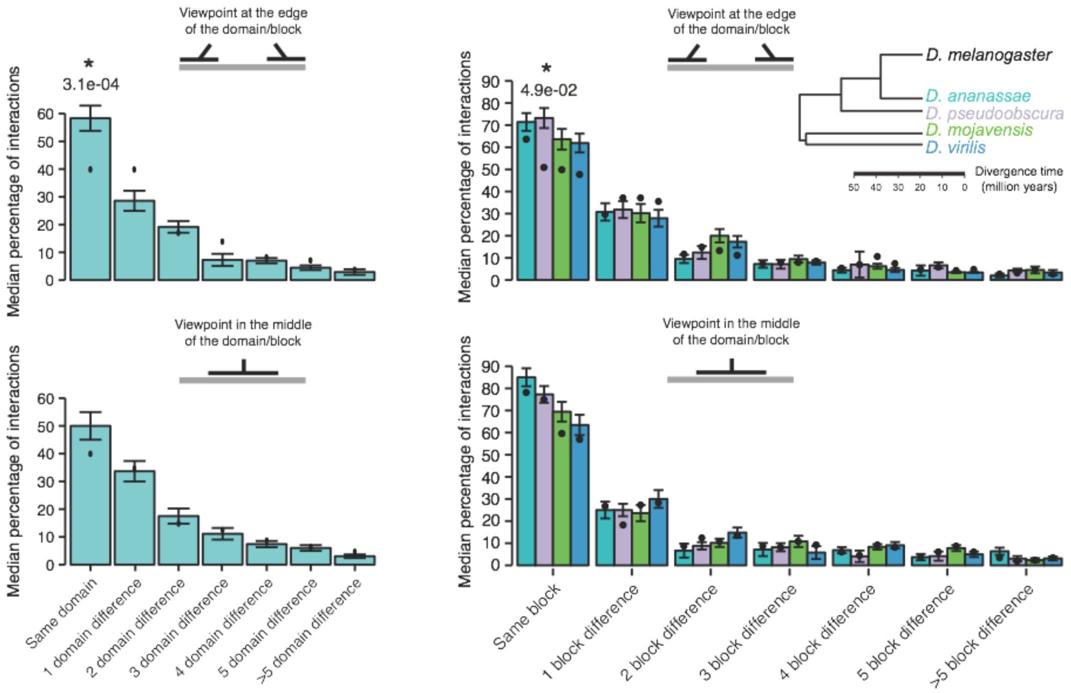

Figure 33. Containment of significant interactions. Measured within the same topological (left-panel), or microsyntenic (right-panel) domains, as the viewpoints they originate from. Viewpoints have been split into edge (upper row), or middle (lower row) groups depending on they relative position within the domain. Bars represent the median percentage, and error bars standard error. Dots are expected percentage of containment as estimated by simulation (see Methods). P-values of comparison between observed and expected percentage have been calculated using two-sided Mann-Whitney U Test.

genomic interactions are likely to be located within the same topologically associated domains (TADs) as the viewpoints (enhancers) themselves I used the previously defined TAD boundaries (Sexton, T. et al. 2012) to calculate the percentage of containment for each of the characterized viewpoint, and compared it to the expected percentage from the simulation (see Methods). Since the location of the viewpoints might influence the percentage of containment (e.g., when the viewpoints are located close to the boundaries of the topological domain, they have to be at least partially asymmetric to constrain the interaction profiles), I divided viewpoints into two categories depending on their relative position within the overlapped domain: edge or middle (see Methods). Astonishingly, the majority (median 60%) of all interactions for edge-viewpoints are found within the same topological domain as the viewpoint itself, which is significantly higher than the 40% expected by chance (Figure 33; P-value = 3.1e-04; Fisher's Exact Test). A similar significant pattern was observed for middle-enhancers, although their difference is somewhat smaller (Figure 33; 50%, compared to the expected 40%). These results are in contrast to when neighbouring TADs were considered, where the median percentage of interactions is equal or lower than the expected level, supporting the hypothesis that although genomic interactions might be distant, they are contained within a specific higher spatial organisation.

Using a similar approach, I also tested the hypothesis that interactions were found predominantly within conserved blocks of syntenty, defined by the sequence comparison between *D. melanogaster* and four other drosophilids -- *D. ananassae*, *D. pseudoobscura*, *D. mojavensis*, and *D. virilis,* spanning an evolutionary divergence time of 50 million years (from Engström, P. G. et al. 2007). Here, interactions tend to be formed within the same block of synteny as the viewpoint itself with an even higher percentage than within Hi-C domains. This is particularly evident when looking between the more evolutionary distant species, such as the *D. melanogaster* and *D. pseudoobscura* in edge-viewpoints, where >70% of interactions are constrained within a syntenic block, while the simulated expectation was only 50% (Figure 33; P-value = 4.9e-03; two-sided Mann-Whitney U-test). Both independent approaches of defining the global regions of 3D organisation, either depending on the unbiased chromosome conformation capture experiment (Hi-C), or conserved microsynteny between drosophilids, present strong and significant evidence that enhancers are highly likely to form genomic interactions within a constrained space, within the boundaries defined by the higher spatial organisation.

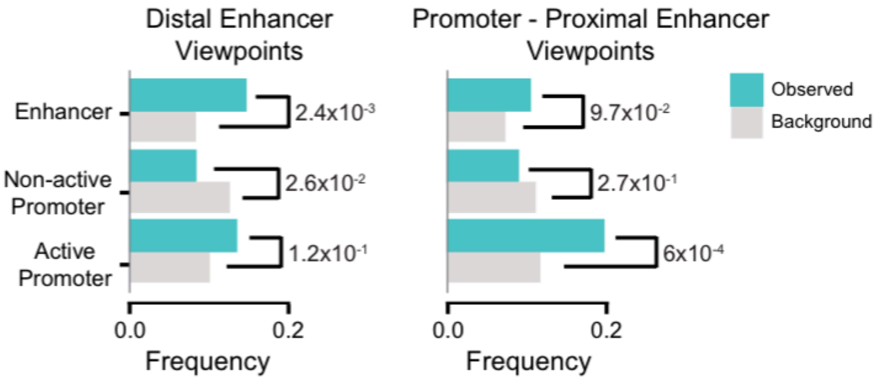

Figure 34. Frequency of overlaps between genomic features and unique significant interactions (1,036). Background proportion has been performed using the same methodology on matched regions (see Methods). Viewpoints have been split into ones within 1kb of TSS (promoter-proximal; 48), and distal enhancers (59). P-values from the comparison of observed, and expected proportion have been calculated using Fisher's Exact Test.

4.2.3 Functionality of interacting regions

4.2.3.1 4C interactions predominantly overlap promoters and enhancers with higher activity

In the previous chapters, I presented evidence that interactions connect various genomic features throughout the genome, while being contained within larger spatial domains. Since genomic elements that share the same functionality and epigenetic state were also shown to be organised in wide genomic regions that can extend for >100kb (Filion, G. J. et al. 2010), I wondered how our developmental enhancers correlate with the levels of gene activity and histone marks. To test this, I constructed a background set of interactions fitted for the GC content, mappability and width, for estimation of the likelihood of particular feature overlap with interactions (see Methods). Distal enhancer viewpoints tend to preferentially contact other enhancers (Figure 34; P-value = 2.3e-03; Fisher's Exact Test), and are depleted for interactions with non-active promoters (see definition in Methods). On the other hand, while still being strongly enriched for contacts with other enhancers and depleted for non-active promoters, interactions of promoter-proximal viewpoints also have a higher than expected likelihood of overlapping other active promoters (P-value = 6e-04;

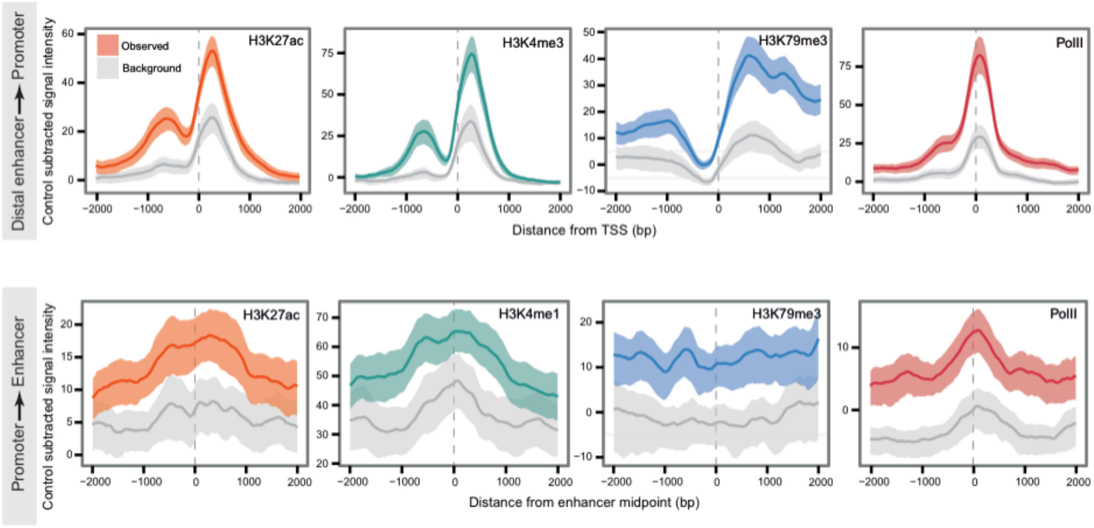

Figure 35. Profiles of histone modifications on 4C interactions. Including marks for active enhancers (H3K27ac; orange), enhancer location (H3K4me1; teal), active promoters (H3K4me3; teal, upper panel), repression (H3K27me3; blue), and RNA Polymerase II occupancy (in red), on promoters (upper row), and enhancers (lower row). In grey are summarized signals of histone modifications on background regions (see Methods). Thick lines are representing the mean trimmed by 10%, and bands 95% confidence interval, as estimating using bootstrapping procedure.

Fisher's Exact Test). Overall, enhancers' interactions are not randomly distributed over the genome, but are preferentially localized with other enhancers and active promoters. In addition to using RNA-Seq, as a readout of gene expression, I also used post-translational modifications of histone H3 tail, epigenetic information that is known to be associated with gene activity (H3K27ac, H3K4me3, H3K79me3), enhancer location (H3K4me1), enhancer activity (H3K27ac, H3K79me3, PolII), and occupancy of the RNA Polymerase II (PolII). Comparable to the pattern observed with RNA-Seq, epigenetic marks representing the activity of genes are significantly increased on interacting promoters of both promoter-proximal and distal viewpoints (Figure 35). More interestingly, interactions between promoter-proximal viewpoints and other enhancers, (which were centered on their midpoints and enriched by the enhancer-specific epigenetic mark H3K4me1), also show a clear increase in histone signals that were found to be predictive of enhancer activity (Figure 35; H3K27ac, PolII, H3K79me3 as shown in Bonn, S. et al. 2012). Overall, our results indicate that while enhancers form a complicated network of interactions, they prefer to contact other genes and enhancers that are highly enriched for activity at the corresponding developmental stage, as inferred from the RNA-Seq and epigenetic ChIP-Seq experiments.

4.2.3.2 Insulator factors occupy significant number of genomic contacts and correlate with higher interaction strength

Genomic interactions are thought to be stabilized through interactions with DNA-associating proteins (*e.g.* the components of the mediator complex), transcription factors, and insulating factors, so it was of great interest to compare insulating factor occupancy with the 4C-interactions defined at overlapping developmental stages. For this, I used ChIP-chip datasets on seven different insulating factors throughout the *Drosophila* genome from 0-12h of embryogenesis (Figure 36; Nègre, N. et al. 2010): GAF, BEAF-32, CP190, SuHw, Mod(mdg4), CTCF_C and CTCF_N. Around 50% of the 4C significant interactions at 6-8h of development overlap with at least one of the insulator binding sites, a percentage that is lower than expected for the single insulator occupancy (Figure 37). This observation is however reversed when combinatorial co-occupancy of four or more insulating factors is compared to significant 4C-interactions (13.3% of interactions, P-value = 10e-27; Fisher's Exact Test). By themselves, 53% of insulator peaks overlap active promoters, and only 5%

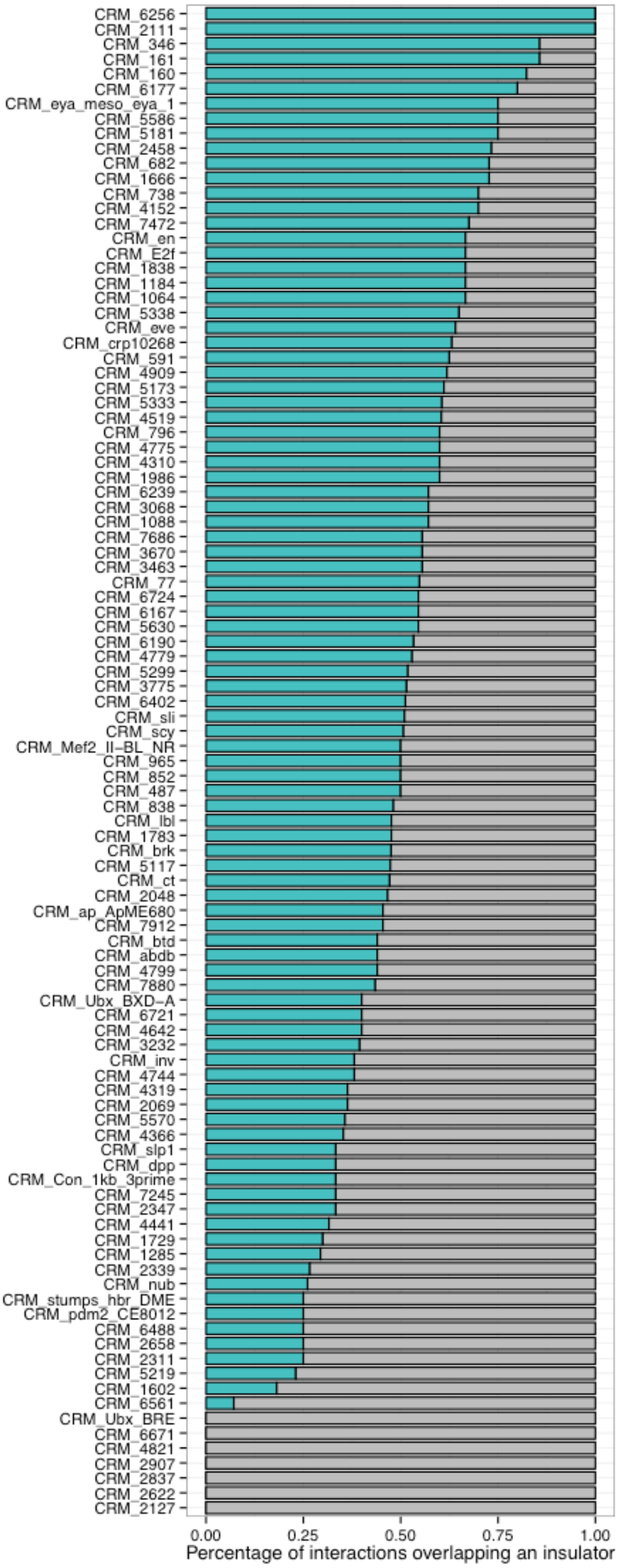

Figure 36. Frequency of insulating proteins on 4C interactions. Distribution of overlap percentages between significant interactions (in teal), and binding sites (defined at threshold of 1% false-discovery rate; from Nègre, N. et al. 2010) seven insulators, including BEAF-32, CP190, Mod(mdg), SuHw, GAF, CTCF_C, and CTCF_N.

with enhancers from our list. This distribution is noticeable altered, however, when considering only those insulator peaks associated with 4C interactions with 16% of insulator-interactions contacting other enhancers, along with the reduced frequency of interacting with active promoters (35%). Interactions that are not bound by any of the seven insulators are extremely depleted for active promoters (1%).

As mentioned above, insulators are statistically enriched for co-occupying the same interacting regions (Figure 39). I investigated this further, by looking at the proportion of every possible co-localization combination. Eight percent of 4C interacting regions are bound by GAF alone, and a significant number of contacts are also linked to various other factor combinations: BEAF_32::CP190 (5%), SuHw alone (4%), BEAF-32 alone (>2%), BEAF-32::CP190::GAF (2%), with all 5 factors (except SuHw, with CTCF merged) being co-bound on 3.7% (37 out of 985) insulator-associated interactions. GAF binding was especially frequent on enhancer contacts, originating from either promoter-proximal of distal viewpoints, while promoter interactions were most significantly bound by the combination of BEAF-32 and CP190. Among viewpoints, there is a continuous range of interactions that are associated with insulator binding, ranging from complete depletion for CRM_Ubx-BRE to 100% overlap with all interactions from viewpoints CRM_6256 and CRM_2111.

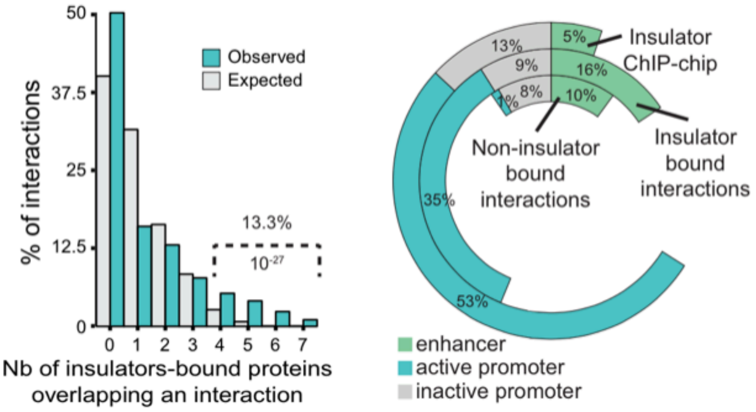

Figure 37. Collective distribution of insulators on significant 4C-interactions. On left, histogram is representing the percentage of interactions overlaps with seven considered insulators, compared to the overlap with the background regions (see Methods). P-value is calculated from the comparison of four or more insulator-bound interactions between observed and expected set (Fisher's Exact Test). On right, distributions of genomic features are shown on insulator peaks (outer layer), overlapped with 4C interactions (middle layer), and non-insulator contacts (inner layer).

Since insulator-bound interactions differ in their feature composition, I also tested a number of other interaction properties including contact frequency (based on both fold change and Z-score) and interaction-to-viewpoint distance that might differentiate insulator and non-insulator bound interactions in individual, grouped and total combinations. Interestingly, insulator-bound interactions tend to have a higher interaction fold change (Figure 38; ratio of observed read counts to the expected fit; 3.4e-03; two-sided Mann-Whitney U Test), but be at the equivalent distance from their associated viewpoints (P-value = 0.35; Two-sided Mann-Whitney U Test). In short, insulators bind a large proportion of genomic interactions defined by our 4C-seq experiment at 6-8h of embryonic development, preferably contacting other enhancers and active promoters. When single peaks are considered, the percentage of overlap was similar to background regions. In contrast, when 4 or more insulators are bound together, they are significantly enriched at interactions, with GAF being the most enriched factor on enhancer-interactions (Figure 40), and combination BEAF-32::CP190 on promoters. Interactions that were bound by insulating factors were found to have a higher read count than expected, albeit at the similar distance from the originating viewpoint as the unbound ones.

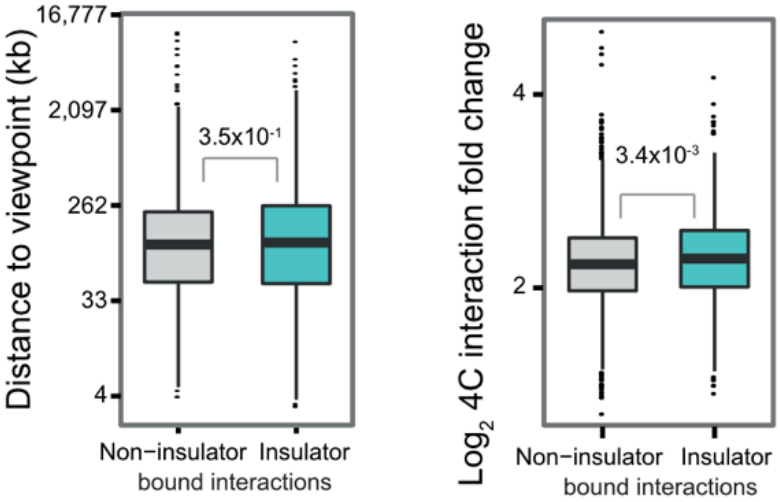

Figure 38. Differences in interaction properties between insulator bound and unbound interactions. Comparison of interactions properties on fragments that are either bound by insulators (in teal), or unbound (in grey). P-values are calculated using two-sided Mann-Whitney U Test.

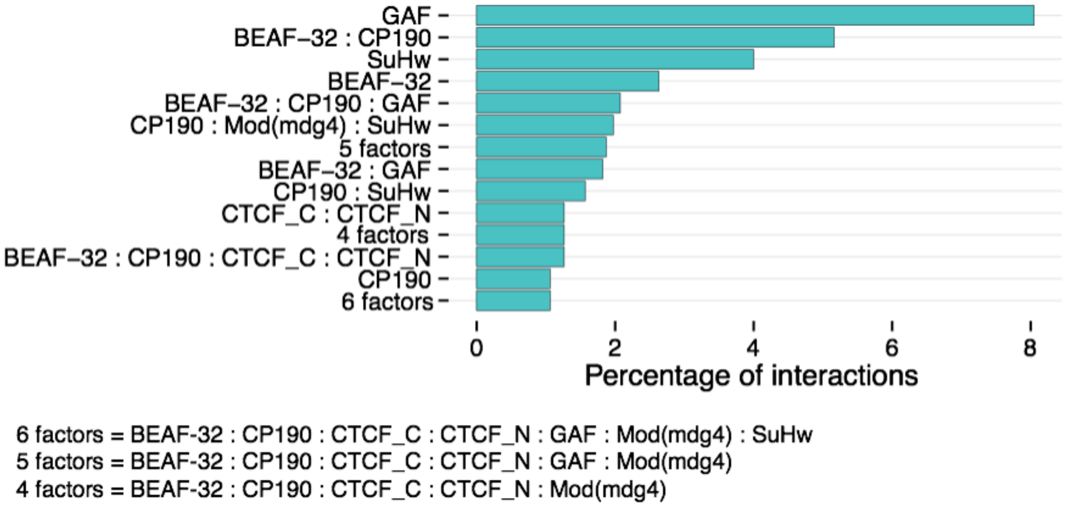

Figure 39. Distribution of insulating protein combinations. Percentage of 4C-interactions overlapped by a single, or combination of insulators.

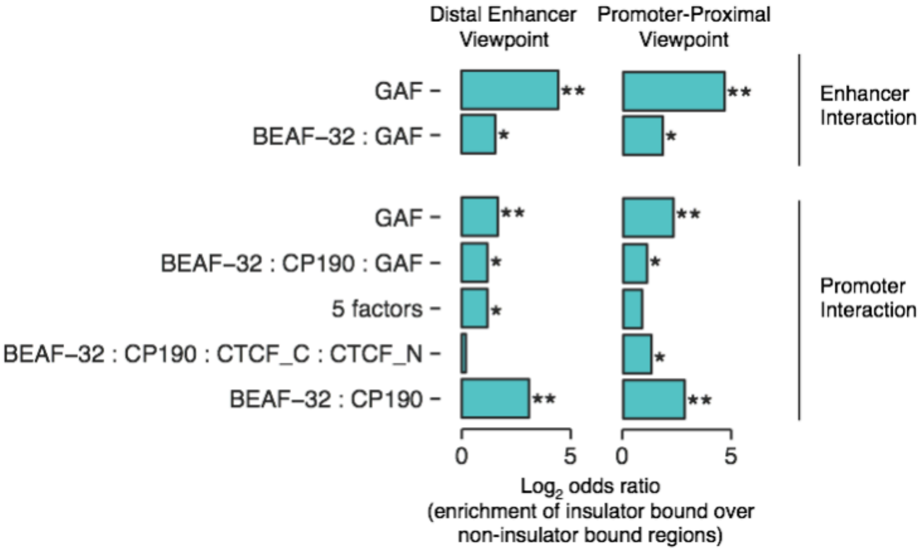

Figure 40. Enrichment of insulating protein combination for 4C interaction overlap. Likelihood of 4C interactions from either distal (left), or promoter-proximal viewpoints (right) to be bound by specific insulator combinations on enhancer and promoter interactions. Enrichment represents the ratio of bound vs. unbound proportion for each category. P-values have been calculated using Fisher's Exact Test (* false-discovery rate <10%; ** false-discovery rate <0.1%).

4.2.4 Stability of contacts between different spatiotemporal contexts

4.2.4.1 4C interaction are surprisingly stable across spatiotemporal contexts

Our experimental setup of two developmental time points (3-4h and 6-8h) and two tissue contexts (whole-embryo and mesoderm) allowed me to explore interaction dynamics in both spatial and temporal directions. Using DESeq2 based on a Gamma-

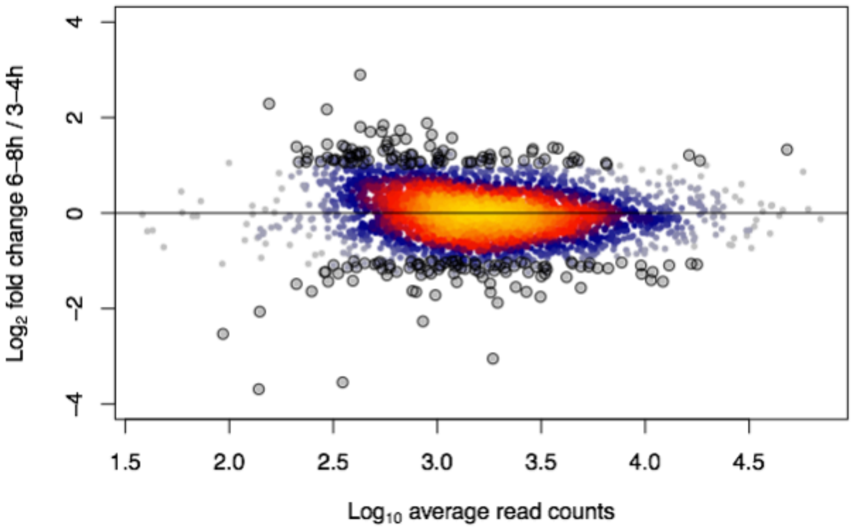

Figure 41. Differential analysis of 4C interactions. Normalized using viewpoint-specific fits, differential plot shows the overall change in interaction strength between the two temporal states: developmental stage when the cells are still multipotent (3-4h), and specified (6-8h). Warmer color gradient indicates the increased point density, and circled dots represent the significant differential interactions at 10% false-discovery rate and 1 Log2 fold change.

Poisson model as a method for differential analyses (Figure 41), I found 177 temporal and 139 spatial significantly changing interactions based on a combined cutoff of fold-change (>1 Log2) and false-discovery rate (<10%). This number represents only 6% of the total interacting fragments, indicating extensive and surprising interaction stability, despite considerable morphological differences between considered contexts, and the fact that up to 30% of all genes are differentially expressed in any given condition (at 10% FDR and 2 (Log2) fold change). A typical example of interaction stability throughout developmental progression can be observed at the *ap* locus, where normalized interaction read counts overlapping the promoter region (originating from ap_ApME680 viewpoint) do not change from 3-4h to 6-8h of development. Despite the lack of significant differences in interaction strength, the expression of the *ap* gene is dramatically transitioning from very few transcripts to high expression level at the later time stage. To further explore the genomic signatures that underlie the described phenomenon, I defined a genome-wide set of 'OFF-to-ON' (459), and 'ON-to-OFF' (277) genes that dramatically switch in their gene expression levels between the matched developmental stages. As in the *ap* example, there is no significant differential fold change difference of 4C interactions in either the off-to-on, or the on-to-off gene set (P-value's of 0.73 and 1, respectively;

two-sided Wilcoxon Test). Overall, differential analysis of 4C interactions revealed that despite the underlying biological differences, there is little spatial reorganization that would follow such change (Figure 43).

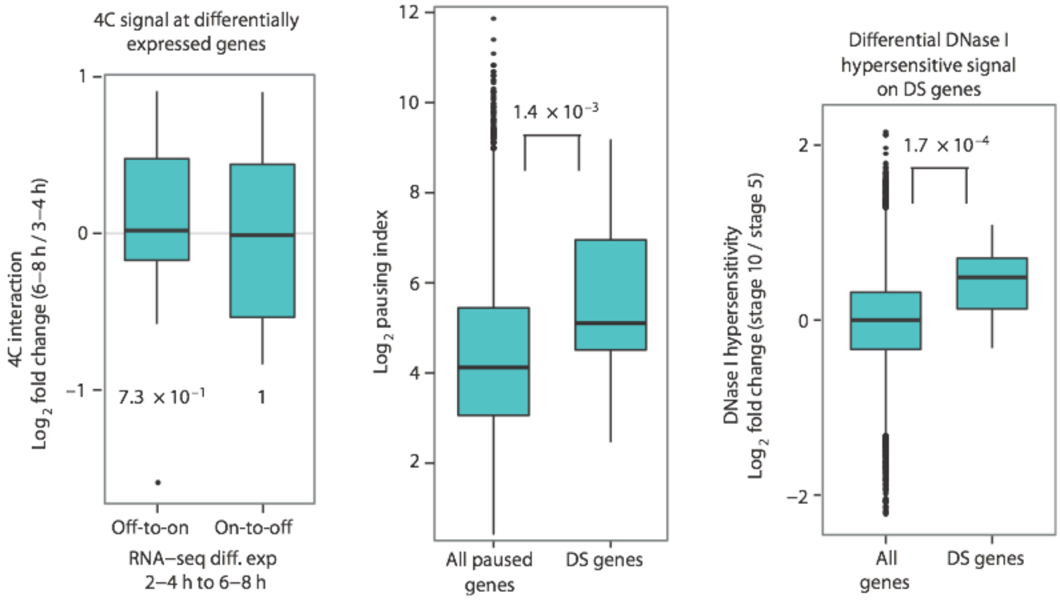

Figure 42. Differential interactions are correlated with paused genes and increased chromatin accessibility. On left, boxplots of 4C interaction log fold change in temporal directions are compared on off-to-on, and on-to-off genes (defined using RNA-Seq; see Methods). Middle panel shows the comparison of pausing index (Saunders, A. et al. 2013), between differentially expressed genes with stable loops (DS genes) and all paused genes. On right, DNaseI hypersensitivity (chromatin accessibility) fold change between the equivalent developmental stages are compared between all, and DS genes.

4.2.5 Association with RNA Polymerase II on DS genes

4.2.5.1 Differentially expressed genes with stable loops (DS genes)

Since the timing of the 4C interactions in our study did not reflect the time when gene expression initiates during *Drosophila* embryogenesis, I wondered if there are any other genomic signatures that would correlate with gene expression. To find such a signal, I integrated the list of 'OFF-to-ON' genes with the stable non-differential promoter interaction, which we called 'DS genes' (differentially expressed genes with stable loops). The lack of gene expression levels at the earlier developmental stages could indicate that these genes are completely inactive (without any RNA PolII preinitiation complex) or that they have paused RNA polymerase (Saunders, A. et al. 2013; Lagha, M. et al. 2013). To distinguish between these two possibilities, I have used 20 strictly defined DS genes and integrated them with data on RNA Polymerase II occupancy (Nègre, N. et al. 2011; Chen, K. et al. 2013) and nascent transcription using GRO-Seq (Saunders, A. et al. 2013).

4.2.5.2 RNA Polymerase II occupancy and GRO-Seq suggest pausing of the DS genes

Accumulation of RNA PolII at the transcription start site, and a high ratio between read counts of nascent transcription on promoters over gene bodies as inferred from GRO-Seq experiments are both indicative of RNA PolII pausing (Core, L. J. et al. 2008; Min, I. M. et al. 2011). Focusing on the TSS regions of DS genes, I summarized read counts representing the occupancy of RNA PolII form a prominent peak at 3-4h, before the onset of transcriptional activity. To confirm that DS genes are indeed paused at 3-4h of embryonic development in *Drosophila melanogaster*, I used three categorizations of pausing ('Top 25%', 'Up to 50%', and all) based on the GRO-Seq experiments at a matched developmental stage (Saunders, A. et al. 2013) in a permutation analysis where the observed pausing percentage among the DS genes was compared to the larger set of 'OFF-to-ON' genes. Permutation tests on DS genes showed a significant enrichment of paused genes (15 out of 18; P-value = 0.022), compared to the 10,000 simulations with shuffled labels (Figure 44). This test was robust to changing the stringency of genes defined as paused (the previously mentioned categories) or to different definitions of the background set, suggesting that the priming (pausing) of RNA PolII is a genomic signature that is strongly associated with the formation of 4C interactions before the onset of transcription. This relationship was additionally confirmed when the levels of pausing index were compared between DS and all other genes (P-value = 0.0014; two-sided Wilcoxon Test). In short, our results suggest that stable 4C interactions, which exist before the gene is transcriptionally active, are highly associated with RNA PolII pausing.

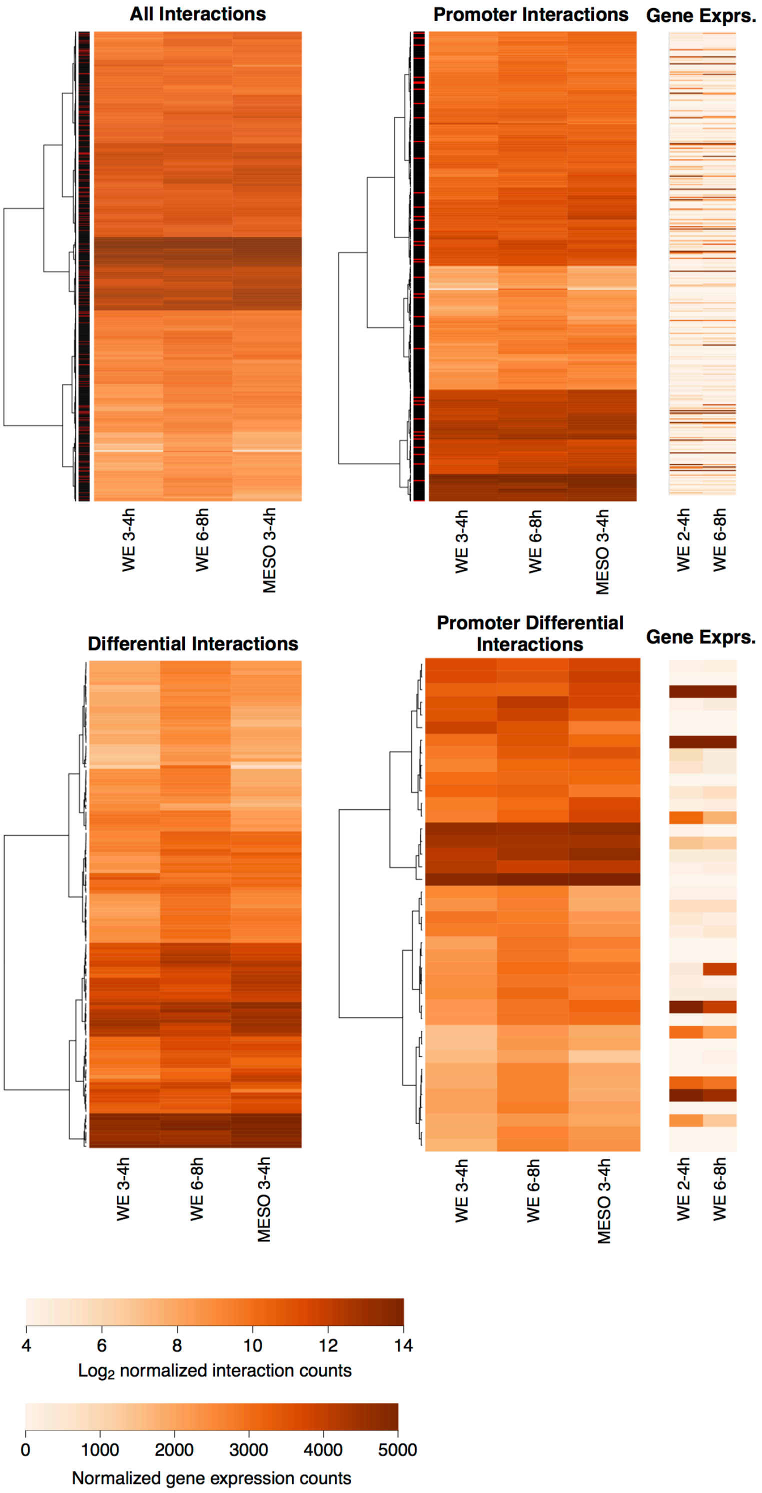

Figure 43. Clustering of normalized 4C-interactions. Counts on shared regions between three experimental conditions: whole-embryo 3-4h and 6-8h, together with the mesoderm-specific contacts at 3-4h (four main panels). Promoter-specific interactions have been matched with the gene expression of the corresponding genes (Graveley, B. R. et al. 2011). Differential interactions have been defined using DESeq2 (see Methods). Heatmaps have been ordered using Euclidean distance and Ward Method. Red lines in the upper panels represent the location of differential interactions.

4.2.5.3 Differential DNaseI hypersenstivity indicates increased TF occupancy on DS genes

Since enhancers regulate the levels and spatio-temporal pattern of gene transcription through the binding of transcription factors (Spitz, F. & Furlong, E. E. M. 2012), the increased binding of enhancing factors at the later developmental stage might explain why DS genes switch on specifically at this stage. To test that hypothesis, I used DNaseI hypersensitivity, which reflects the binding of a range of transcription factors on the interactions overlapping the DS genes (Degner, J. F. et al. 2012; He, H. H. et al. 2014). Differential analysis of DNaseI data revealed a significant increase in chromatin accessibility at 6-8h of development compared to the whole-genome average (Figure 42; (P-value = 0.0017; two-sided Wilcoxon Test), supporting the model that some of the enhancer-to-promoter interactions are preformed prior to the onset of transcription and primed with RNA Polymerase II (Figure 45). This may act to prefigure developmental genes for rapid activation, with stage- and tissue-specific activation finally triggered by the binding of specific transcription factors that release the paused polymerase, causing a significant increase in gene expression.

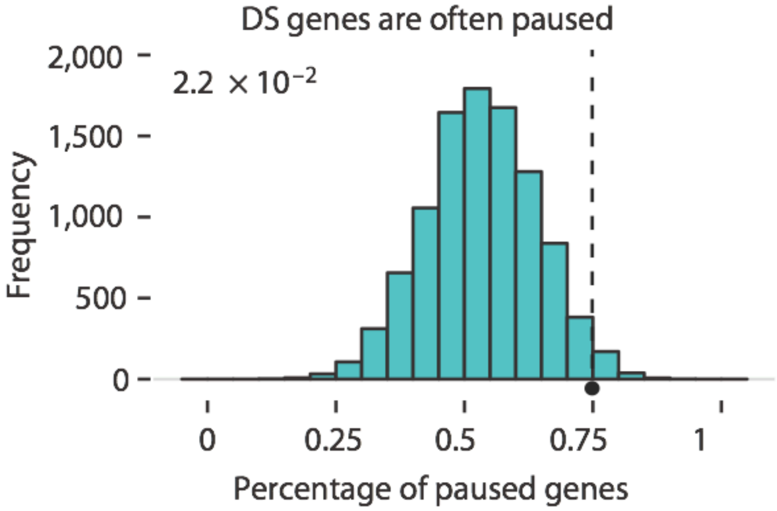

Figure 44. Differentially expressed genes with stable loops (DS genes) are significantly enriched for pausing category. Estimated using 10,000 permutations, percentage of paused genes between DS genes (18) were compared to all off-to-on genes (459). Dotted line represents the observed, while histogram shows the distribution of expected pausing percentages. P-value is shown from the permutation run.

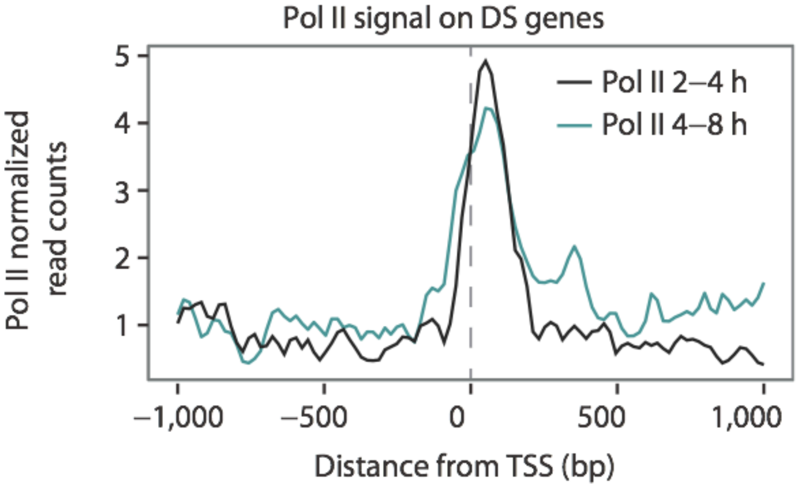

Figure 45. Profile of RNA Polymerase II occupancy on DS genes. From two stages of development (2-4h in black, and 6-8h in teal). Quantile normalized read counts were summarised using 10% trimmed mean around transcription start sites of DS genes.

4.2.6 Summary

My initial findings, that the repressive Polycomb system is often recruited to developmental enhancers, highlights the complexity of developmental enhancers and how they can function in a context-specific manner. In light of this result, I expanded my research interest to investigating the consequence of binding of other factors (e.g. insulator proteins) and epigenetic states on enhancer regulation. These analyses, however, are complicated by the simple fact that even within the compact *Drosophila* genome, regulatory elements can be scattered far away from their target promoters. As part of the lab's efforts to bridge this gap, I collaborated on a project describing enhancer-target interactons, as described in this chapter.

In the biggest 4C-seq study up-to-date, I characterized the extent of three-dimensional interactions for more than 100 different viewpoints in two temporal (3-4h, and 6-8h of embryogenesis) and spatial (whole-embryo and mesoderm) contexts. Following the experimental work by Yad Ghavi-Helm, and initial quality assessment that involved optimizing the statistical methods based on 10 control cases by Felix Klein, I defined and analysed the biology of thousands of new interactions originating from 48 promoter-proximal, and 59 distal enhancer viewpoints. Our results discovered three major findings: 1. A surprising number of spatial contacts spanning very large genomic distances, with more than 50% of 4C-interactions being longer than 100kb, with some (e.g. unc-5 locus) reaching even up to 0.5Mb. Such long-range regulation is typically associated in mammalian organisms. There were some long-range interactions previously identified in *Drosophila*, but these mainly occur in the context of Polycomb mediated repression. Our data shows that long-range enhancers interactions associated with active transcription are also very prevalent within the *Drosophila* genome; 2. Even though *Drosophila* embryos undergo significant developmental changes, with cells shifting from multipotency to specification during the selected developmental stages (with 30% of genes being differentially expressed), interactions seem to be remarkably stable, and are often formed even before the onset of transcription.; 3. Examining the list of DS genes, genes defined by shifts from very low expression at 3-4h to a very high transcriptional activity at 6-8h of development while keeping the same interaction strength, revealed prominent RNA Polymerase II occupancy at the early time stage, suggesting that enhancers and paused promoters are spatially close prior to the onset of transcription. Such relation might plausibly exist to allow the sudden release of RNA Polymerase II and burst of transcription in specific spatiotemporal context, as suggested by the differential DNaseI hypersensitivity analysis that revealed an increase of transcription factor occupancy on DS genes at the later stage of development.

4.3 High-resolution Hi-C chromatin interactions

4.3.1 Definition of significant enhancer and promoter contacts

4.3.1.1 Extension of the 4C-seq study

Recent advances in sequencing and chromosome conformation capture technologies (Lieberman-Aiden, E. et al. 2009) allowed me to expand our previous research on the enhancer interactome from 100 viewpoints used in the 4C-seq study to all enhancers in the *Drosophila* genome, up to an estimated number of ~40,000 elements (Kvon, E. Z. et al. 2014) using high-resolution Hi-C experiments (Figure 46), which were performed at two developmental time stages (3-4h and 6-8h) in two biological replicates. The data consists of at least 5 technical replicates resulting in total of 6.5 billion sequenced reads, resulting in the most detailed interactome study in *Drosophila melanogaster* to date. Focusing the research on intrachromosomal one-to-one (*cis*) read pairs, I could now address questions involving the full complexity of enhancer-promoter interactions, and their relationship to transcription factor occupancy, epigenetic state and developmental context, and integrate this information with additional genome-wide information on the levels of transcription, gene functionality and chromatin accessibility to address the following questions: What is the consequence of differing combinations of transcription factor enhancer occupancy and their changes throughout development? How do short and long-range interactions differ in terms of the features and functions of their interacting regions? Are there differences in the number of interacting enhancers depending on their functional categorization? We are also collaborating with Giacomo Cavalli's group, who performed a similar high-resolution Hi-C experiment, but at a much later developmental time point (16-18h; Sexton, T. et al. 2012). This allowed me to additionally explore the interaction dynamics between three distinct developmental stages.

The Hi-C data in the Furlong Group was generated through collaboration with postdoctorate member Yad Ghavi-Helm, Ph.D., who collected the staged *Drosophila melanogaster* embryos, performed Hi-C experiments in the wet-lab environment, prepared the chromatin and the Illumnia sequencing libraries, which were sequenced at EMBL Genomics Core Facility.

4.3.1.2 Reproducibility and modeling of interactions

Biological replicates show high levels of reproducibility (ranging from Pearson's correlation coefficient of 0.55 for 6-8h, and 0.81 for the 3-4h samples; Figure 47), indicating the high quality of these datasets. To increase sensitivity, all samples were merged into a single contact matrix to increase the sensitivity of detecting the interacting regions, followed by filtration steps which included the removal of observed biases such as the large increase of read count frequency for particular orientation (inward and outward) of paired-end reads within 10kb distance and equal-fragment mappings. P-values for each individual anchor-to-fragment pairs were estimated using a Negative-Binomial model through the comparison of observed and estimated expected contact frequency (see Methods). To define the significant interactions, I used a stringent threshold requirements based on a comparison to 4C-seq anchors of P-value lower than 5%, with at least 10 supporting read counts, which roughly corresponds to a specificity of 90.72%, and sensitivity of 22.01% (see Methods; Figure 10).

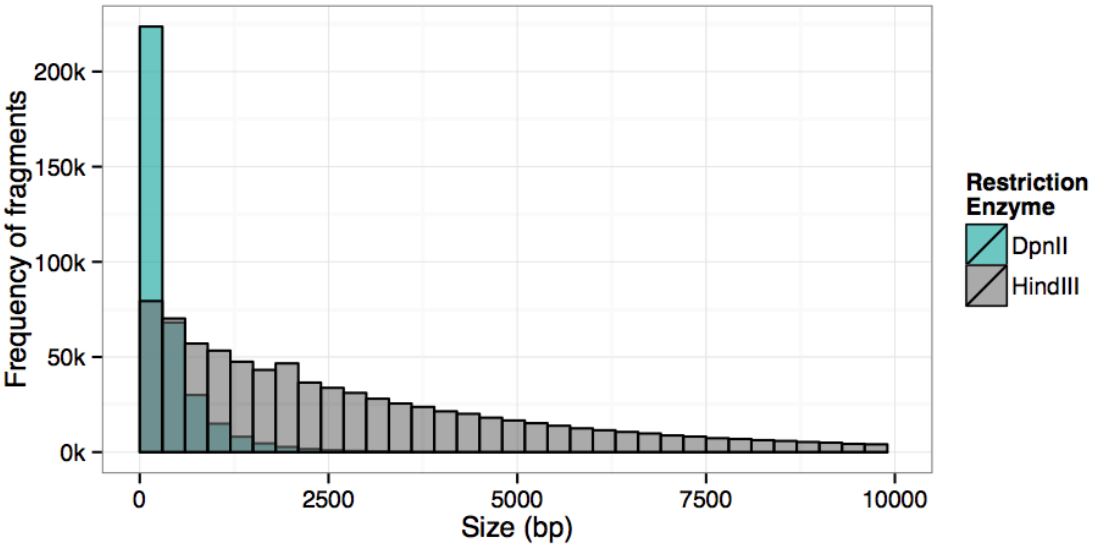

Figure 46. Distribution of theoretical DpnII fragment sizes. Using the same restriction enzyme (DpnII) as in the 4C-seq project (Ghavi-Helm, Y. et al. 2014), Hi-C experiments result in both comparable anchors-viewpoint fragments, and higher resolution (~350,000 fragments, median length of 193bp) compared to HindIII used in human cell line Hi-C study (~840,000 fragments with median width of 2.27kb; Jin, F. et al. 2013).

4.3.1.3 Quantification and properties of significant contacts

There are, in total, 22,629 significant interactions that originate from a focused anchor space of 4,773 mRNA-producing promoters, and 1,972 enhancer regions (7,502 interactions). Additionally, I split the promoter anchors into two activity categories -- 'Active', and 'Non-Active' (7,733 and 7,394 interactions, respectively) based on gene expression levels at 3-4h or 6-8h of embryonic development as estimated using RNA-Seq (see Methods 3.2). On average, there are 2 interactions per anchors (across all three anchor types: active promoters, non-active promoters and enhancers) with an overall median interacting regions width of 998 bp..

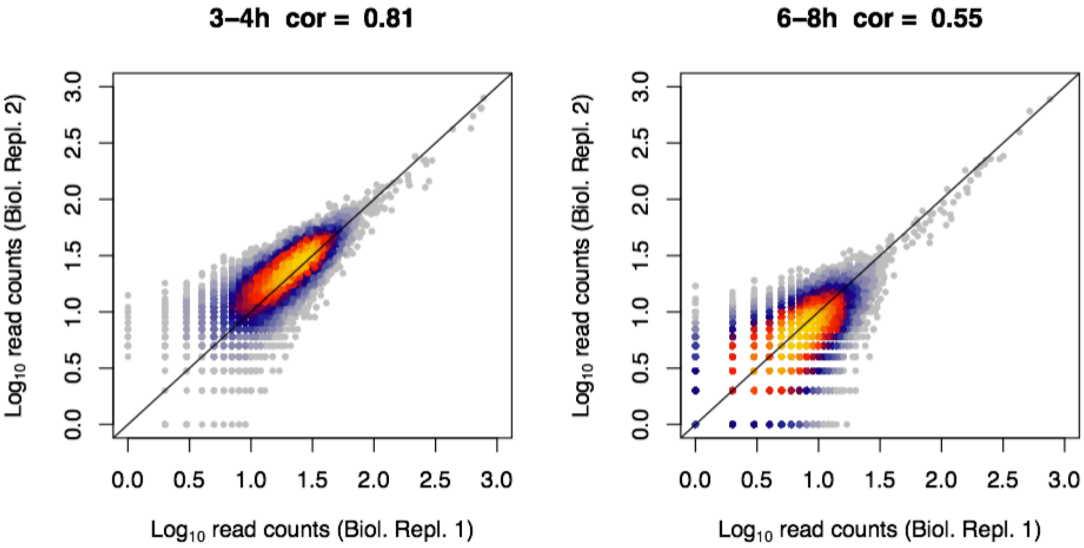

Figure 47. Reproducibility of biological replicates used in Hi-C project. Read counts were summarised on significant interactions (see Methods), and compared using Pearson's correlation coefficient.

4.3.2 Sensitivity of long-distance Hi-C interactions

One of the major findings of our 4C-seq study was an unexpectedly long distance between viewpoint-to-interactions, with and average of 110kb, and some reaching even up to 0.5Mb (Ghavi-Helm, Y. et al. 2014). Given the lower overall sensitivity of Hi-C compared to 4C, for the anchors assessed, I observed a significantly lower overall proportion of very long distance interactions above 100kb (Figure 48; 12.17%), with the median HI-C length ranging from 20.35kb and 33.86kb (active promoters and enhancers), to 60.78kb (non-active anchors). Although 4C-seq and Hi-C experiments were performed by the same person, at the same stages of embryogenesis in the same genetic strain using the equivalent restriction enzyme, there are two key technical differences which likely explain the apparent difference in the observed contact distance: 1. The 4C-seq study was designed on a specific set of 100 distal and promoter-proximal enhancers, and despite best efforts to sample a wide range of activity state and chromatin contexts, 100 viewpoints still represents less than 1% of the estimated number of enhancers in *D.melanogaster* genome (Kvon, E. Z. et al. 2014); 2. More importantly, despite the unprecedented sequencing coverage of our Hi-C study, there is a significant dilution of the information content from the focus on a single viewpoint (4C one-to-many) to 350,000 anchors. For example, a viewpoint based on the known enhancer CRM_Mef2_II-BL_NR contains in total 1,459,239 read counts in our 4C-study, while the equivalent DpnII fragment in our deeply sequenced Hi-C matrix contains only 1,877 interacting reads. As the interaction frequency is inversely proportional to the linear genomic distance from a view-point, long-range interactions will by definition have lower interaction frequency and therefore less reads, making their detection very sensitive to the number of reads obtained in Hi-C experiments.

To prove that the major cause of the reduction in long-range interactions above 100kb in the Hi-C dataset relative to our 4C study is due to a significant increase in the potential interaction space rather than a discrepancy between the methods (either experimental or statistical), I compared three 4C-seq derived summary statistics (Figure 49; viewpoint-normalised read count score, z-score, and interaction fold change) to the Hi-C ones (observed read counts, expectation values, and p-value from the negative binomial test), on 1,960 WE 6-8h 4C-defined interactions (if there were no Hi-C read counts, P-value was set to 1) in two groups of inferred interactions -- lower than 100kb, and higher than 100kb (referring to the interaction distance to the corresponding viewpoint). For both the short, and long distance group 4C-seq interactions show similar levels of z-scores, while having even higher fold change of the read counts compared to the expectation (IFC) for the long category, although there is a noticeable decrease in overall read counts support longer interactions (normalized score). The Hi-C interactions display the same trend of lower observed read counts for longer interactions, albeit with very different consequence. Since very few read counts support any of the interacting regions from the 4C-seq study (positive test cases), it causes a sharp increase in the estimated statistical significance compared to the expected read counts (P-value = 1.7e-53; Mann-Whitney U Test). Strikingly, 415 out of 1,068 defined 4C-interactions that are more than 100kb away from their viewpoints show a complete lack of Hi-C reads, in contrast to 16 out of 892 4C interactions within 100kb of their viewpoint. This indicates that the extremely large anchor-to-contact space causes a lack of power to detect the long distance interactions. Despite this reduced sensitivity of the Hi-C data to detect interactions >100kb away, the data considerably increases the number of long-range interactions known within the *Drosophila* genome.

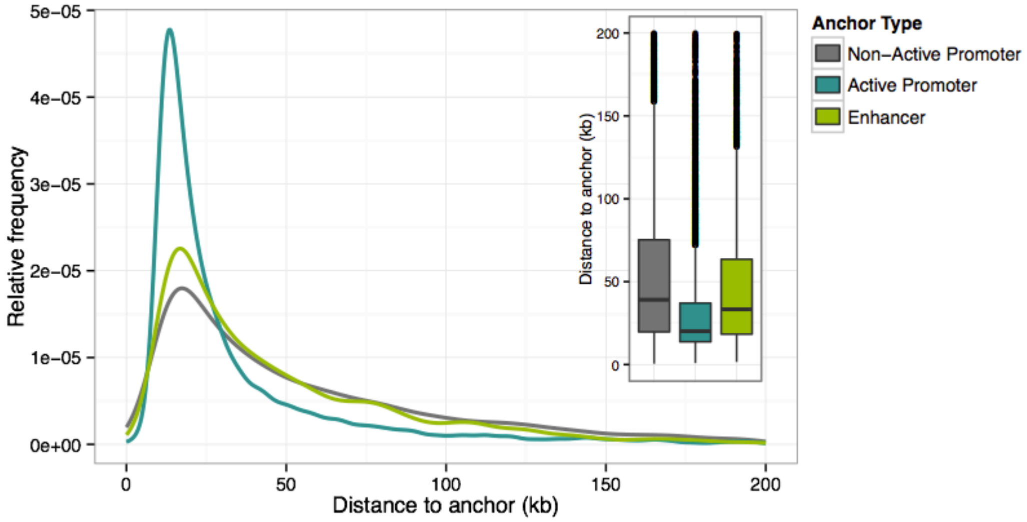

Figure 48. Distribution of anchor-to-interaction distances for three types of promoter and enhancer anchors.

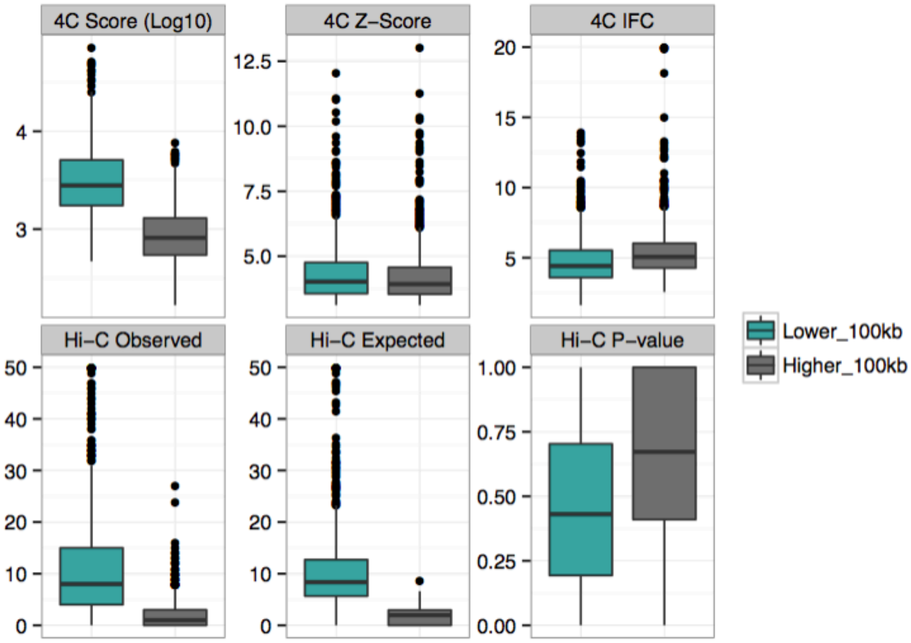

Figure 49. Comparison of 4C-seq and Hi-C interaction values. 4C values (normalized score, z-score in comparison to viewpoint-specific fit, and interaction fold change) to Hi-C observed read counts, expected scores and calculated P-value summarised on the significant interaction regions as defined in 4C-seq project, and divided into shorter (teal), and longer (grey) distance groups.

4.3.3 Functional description of Hi-C interactions

4.3.3.1 Activity of Hi-C interactions correlates with the activity of their originating anchors

Following the determination of thousands of novel genomic interactions from Hi-C data, I have used annotations to explore how do promoter and enhancers anchors differ in their distribution of genomic features. To do that, I assigned each individual interaction to previously annotated genomic elements (promoters, enhancers, intergenic and intragenic regions; Figure 50). Interestingly, a high proportion of active-promoter anchored interactions (21%) show evidence of interacting with other active promoters. A similar pattern can be observed with enhancer-to-enhancer (14%), and non-active to non-active promoter loops (12%), confirming previous findings that functionally similar elements in the genome tend to be spatially localized. This pattern of similarity association is also visible from the K-mean clustering of the representative histone marks (H3K27ac, H3K4me3, H3K36me3 and H3K79me3 along with Pol II occupancy and chromatin accessibility from H3 signal; Figure 54). These marks are strongly correlated with functional activity: Active promoter anchors are significantly enriched in active histone marks (H3K4me3, H3K27ac, H3K36me3 and H3K79me3; Figure 54 upper-left panel), a trend opposite of non-active promoters, showing the clear correlation between RNA-Seq defined classification and observed epigenetic state. Enhancer anchors (upper-right panel) display a wider range of states, having both a repressive H3K27me3 and enhancing H3K27ac signals alongside H3K4me1, a general mark of regulatory elements. Clustering of the histone signals on interactions reveals a more complex combination of epigenetic states compared to anchors, albeit still being reflective of anchor activity. While the majority of active anchors interact with regions of high transcriptional activity and low nucleosome occupancy, a small proportion of active anchors interact with regions with marks indicative of both repression and enhancer locations, suggesting contacts with Polycomb elements. Enhancer-based anchors on the other hand are clearly depauperate of promoter mark H3K4me3, suggesting that most of the interactions are located on the other enhancers, in either repressive (H3K27me3) or active state (H3K27ac).

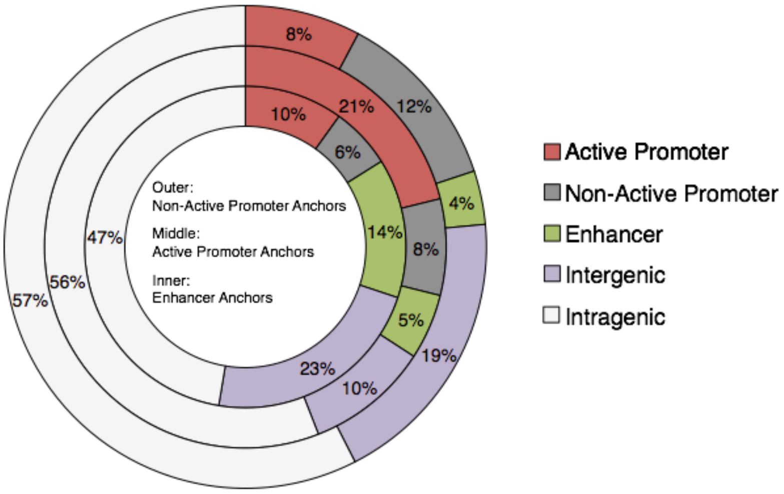

Figure 50. Distribution of genomic elements overlapping significant Hi-C interactions. Split for Non-Active Promoter (outer layer), Active Promoter (middle layer), and Enhancer anchors (inner layer).

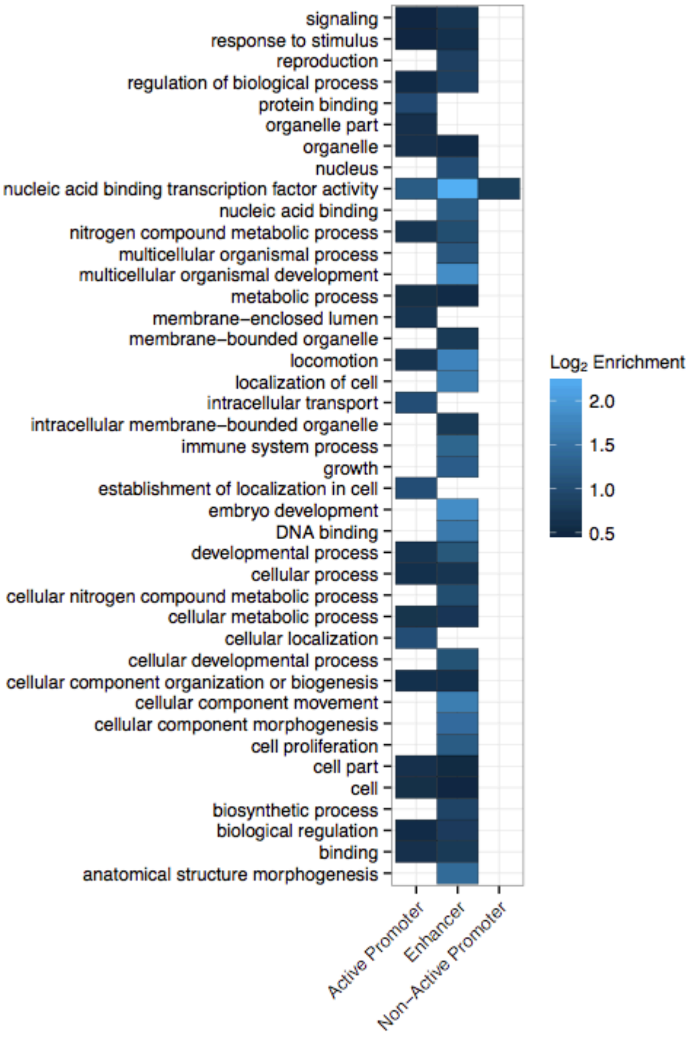

Figure 51. Gene ontology analysis of promoter interactions originating from three different types of anchors. Enrichments represents the odds ratio (strength of association). Only statistically significant terms (false-discovery rate lower than 10%) from all three parts of ontological analysis (molecular function, cellular components and biological process) are shown, and blank otherwise.

4.3.3.2 Hi-C interactions originating from enhancer anchors are significantly enriched in wider range of biological functions compared to anchors of other types

Since Hi-C interactions originate from anchors of differing activity (based on RNA-Seq), and TF occupancy, I explored whether such variety is also reflected on the type of genomic functions of genes they contact. To do that, I used gene ontology analysis (focusing only on the promoter interactions) with the 'Parent-Child-Union' relationship (see Methods) that revelead the clear difference in biological functionality of the targeted genes dependent on the source of the interaction. While non-active promoter anchors exhibit almost no significant enrichment (Figure 51, third column), which is most likely due to the inactivity of most of their promoter contacts at the measured developmental stage, active promoters are contacting other promoters associated with genes involved in basic cellular and developmental processes as well as metabolic activity. On the other hand enhancer-to-promoter interactions are significantly enriched for more specific functions such as multicellular development, morphogenesis, proliferation, as well as immune system process and growth, with the highest log odds ratio for transcription factor binding activity (2.26), suggesting that these interactions are specifically enriched for genes involved in crucial biological processes involved in the proper development of the *Drosophila* embryo.

4.3.3.3 Genes with distinct biological function form differing number of Hi-C interactions

To explore how the basic Hi-C interactions properties differ among anchors with distinct biological functions, I categorized anchor promoter based on FlyBase biological processes (Figure 52; St Pierre, S. E. et al. 2013). Interestingly, genes associated with locomotion (including both cellular and whole-organism movements) and pigmentation (containing genes such as eyes absent, *eya*, involved in mesoderm and ventral cord development; Vining et al. 2005, Xiong et al. 2009) tend to have the highest number of significant contacts (4.05, and 3.95 respectively). In contrast, apoptosis associated anchors, although having the lowest average frequency (2.86), often interact with enhancers (mean of 0.29), which within the context of mid-embryogenesis is probably indicative of contacts with genes responsible for tissue formation as for example the ones adjacent to the segment border (Rusconi, J. C. et al.

2000). Most anchors have remarkably similar interaction distances (Figure 52), apart from pigmentation-labeled ones, which is partially due to high proportion of interactions with intergenic regions (highest among categories at 25%).

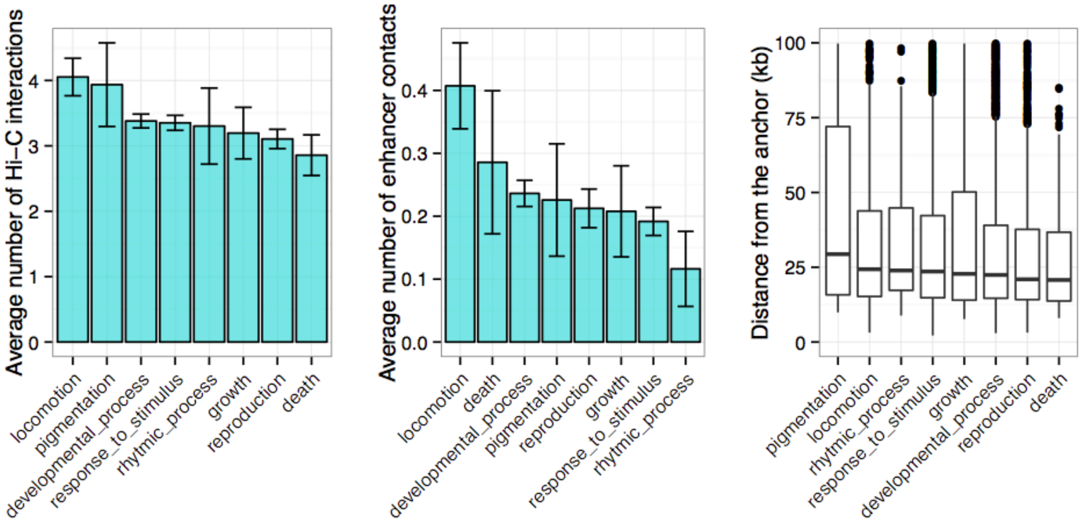

Figure 52. Properties of Hi-C interactions with varying biological functions. Contact frequency summarised by mean (left panel), enhancer contact frequency summarised by mean (middle panel), and distance from anchor midpoint (right panel) for promoter anchors grouped according to the gene ontology biological process categories. Groups are ordered according to the decreasing average value. Error bars represent standard error.

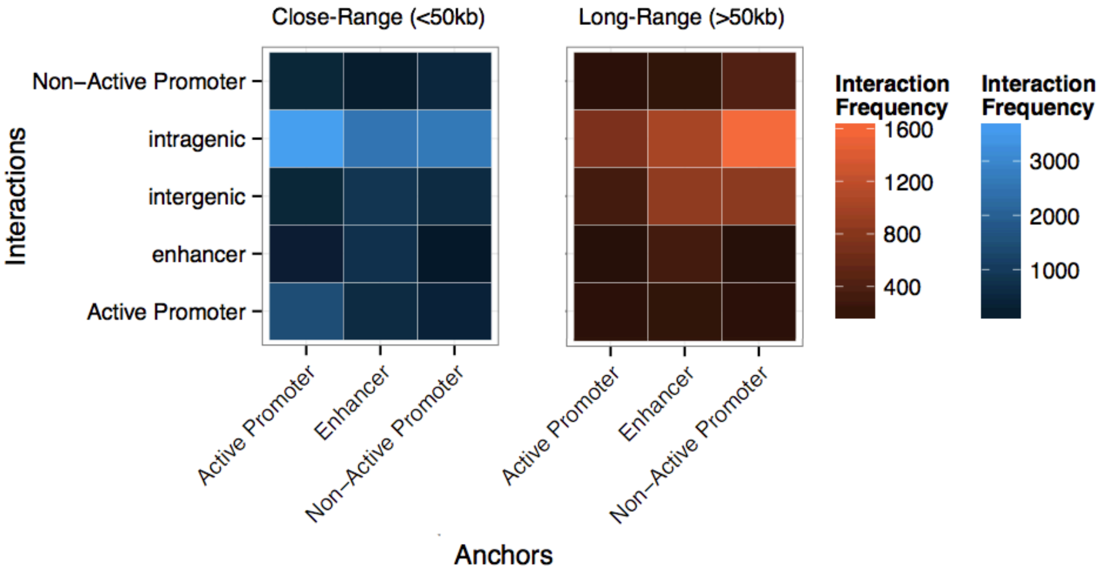

Figure 53. Comparison of short and long-range Hi-C interactions. Interaction frequency for close-range (within 50kb of their originating anchors; blue gradient), and long-range anchors (further than 50kb; orange gradient) split according to the anchor types, and genomic features that overlap the interactions.

4.3.3.4 Comparison of short and long range interactions

Compared with vertebrate systems, the *Drosophila melanogaster* genome is relatively compact (detailed in 4C-seq results chapter) with only ~135Mb mappable nucleotides (euchromatin), despite similar numbers of mRNA-coding genes (~16,000), and isoforms (27,525). Due to such compaction, some enhancers might have to skip several promoters to reach their target gene, while others might act in proximity. To investigate whether there are differences in contact frequency between short (<50 kb) and long range interactions (>50kb), I have split the Hi-C data into interactions acting at an anchor-to-target distance of <50kb (referred hereafter as short) and >50kb (referred to as long). Anchors of all three types with shorter loops primarily interact with intragenic regions, but also include a noticeable number of active-to-active promoter interactions (Figure 53). Longer-range interactions visibly differ in their contact profile, with more associations with intergenic loci, as well as an increase in overall contact frequency from non-active promoters. Enhancer anchors have no striking difference between long and short range, indicating the rather uniform distance distribution of enhancer-to-promoter interactions.

4.3.4 Dynamics of Hi-C interactions throughout embryonic development

One of the major findings of our 4C-seq study was the remarkable stability of the interaction profiles from both promoter-proximal and distal viewpoints; only ~6% of contacts showed significant differences between conditions, despite considerable spatiotemporal changes in gene expression and consequentially morphological phenotypes. I extended the same type of differential contact analysis to all promoter and enhancer anchors using the extended HiC dataset.

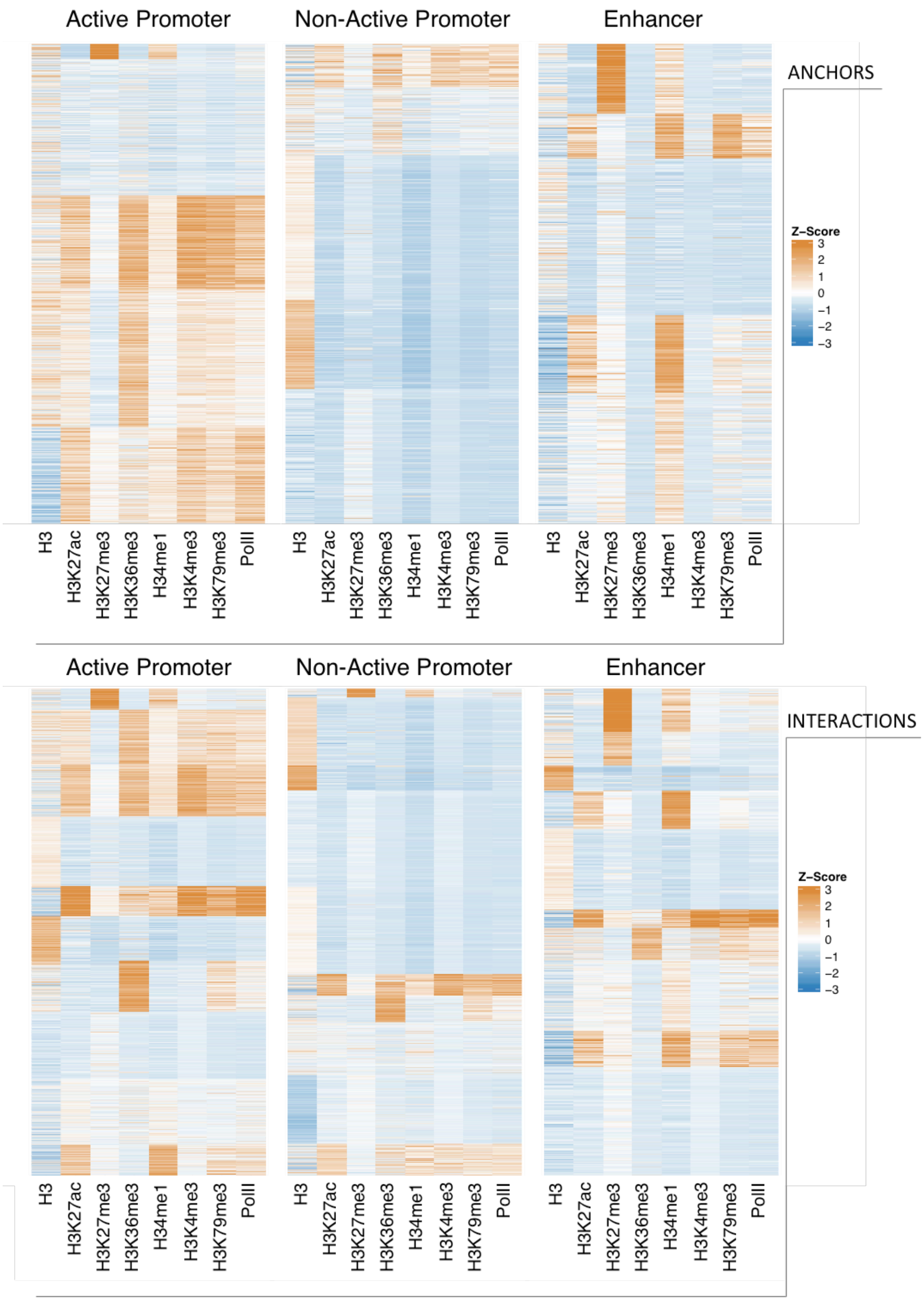

Figure 54. Epigenetic signature of Hi-C anchors and interactions. Epigenetic state of Active Promoter (left panels), Non-Active Promoter (middle), and Enhancer anchors (right panels) summarised using mean statistic in +/- 2kb window around DpnII-fragment midpoints, and converted to Z-score values. Shown are nucleosome occupancy (H3), histone marks representing active and repressed enhancers and promoters (H3K27ac, H3K27me3), exon/intron usage (H3K36me3), enhancer locations (H3K4me1), promoter activity and location (H3K4me3), active transcription (H3K79me3), and RNA Polymerase II occupancy (PolII).

4.3.4.1 Larger differences in Hi-C interaction strength are observed between more distant stages of embryogenesis

In the whole-embryo comparison of developmental progression from multipotency (3-4h) to specified cell state (6-8h; Figure 55) there are 96 significant differential interactions at a 10% false-discovery rate and absolute fold change higher than 2, which is only 0.42% of the total number of tested regions. As it is not feasibility to reliably estimate the dispersion for single replicate experimental designs (as in the Cavalli's 16-18hr data), a comparison of our data to the Cavalli group's was limited to an approximation of differential contacts from the observed fold changes. However, even this simple analysis reveals a 3.7x increase (from 1,017 to 3,723) in the number of regions with fold change higher or equal to 2, indicating that the 3D interactions increasingly differ as the developmental gap widens (albeit this being only an estimation due to the afore mentioned lack of biological replicates that are required for variance estimations).

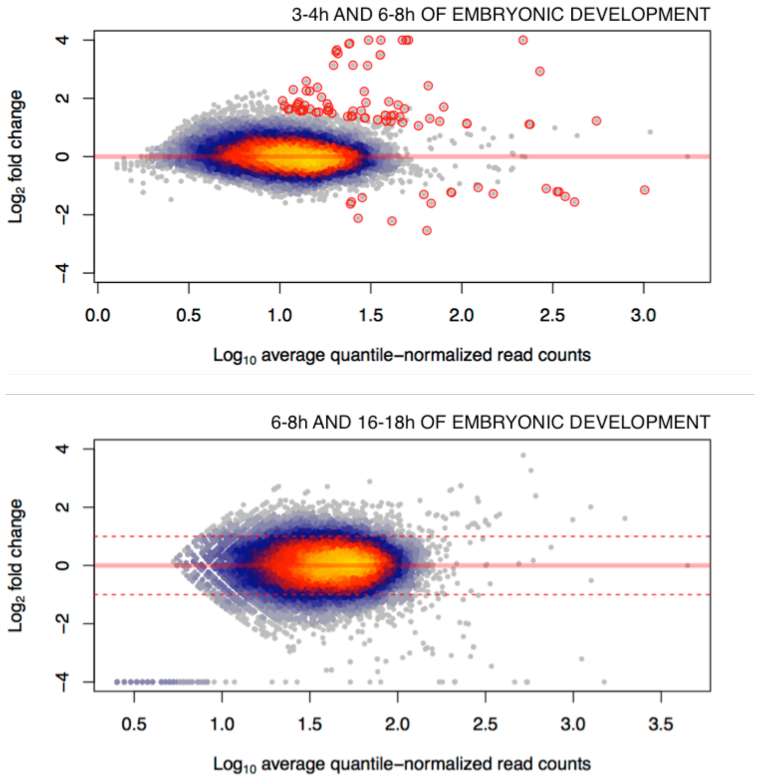

Figure 55. Dynamic of Hi-C interactions. Top panel represents the differential comparison of the progression from multipotency (3-4h) to specified cell state (6-8h). Significant fragments (less than 10% false-discovery rate and 2 fold change) are labeled by a red circle. Lower panel is the approximation of interaction strength changes between more distant developmental stages (sample from 6-8h, in comparison to Cavalli Group's 16-18h sample).

4.3.4.2 Change in TF occupancy on enhancer anchors

Enhancer regions used as anchors in this study are predominantly defined from ChIP-chip occupancy data for five mesodermal transcription factors - Twist, Myocyte enhancer factor-2 (Mef2), Tinman, Biniou, and Bagpipe, three of which (Twi, Mef2, and Tin) have temporal information from equivalent Hi-C stages (3-4h, and 6-8h of embryonic development; Figure 56). To define several global states based on quantitative signal of TF binding, I have used K-means clustering method. Set to define 5 discrete groups for these three factors, clustering revealed distinct dynamic occupancy of Twi (both oncoming and releasing at 6-8h; 382 and 625 interactions respectively), and Mef2 (oncoming at 6-8h; 655 contacts), along with low-bound and continuous occupancy groups (2,158 and 364). I compared the dynamic enhancer states to the dynamics in their interacting promoters, using four temporal datasets : (1) whole-embryo RNA-Seq representing quantitative levels of gene expression (Brown, J. B. et al. 2014, Daines, B. et al. 2011), (2) whole-embryo DNaseI hypersensitivity measuring the levels of chromatin accessibility (including the occupancy of various transcription factors; Thomas, S. et al. 2011) from stages 5 and 10 (corresponding to 3-4h, and 6-8h of development), (3) quantile-normalised whole-embryo RNA Polymerase II occupancy (Nègre, N. et al. 2011), and (4) the interaction change itself (from the dynamic analysis above). Continuous binding of more than one mesodermal transcription factor on the enhancer is correlated with the highest increase in gene expression at a later time stage, with a significant median deviation from 0 (Figure 57; P-value = 1.7e-04; one-sample Wilcoxon Test), but not with the overall TF occupancy, interaction strength nor RNA Polymerase II binding, unlike oncoming of the Mef2 occupancy which correlates with an increase of PolII signal on the targeted promoters (P-value = 8e-05; one-sample Wilcoxon Test).

The largest (2,158 interactions), and perhaps most intriguing class of enhancer states are low-bound by mesodermal transcription factors (grey cluster from above analysis), which in addition to significantly correlating with an increase in transcription (P-value = 1e-05; one-sample Wilcoxon Test) and PolII occupancy (P-value = 1e-04; one-sample Wilcoxon Test), also have greater interaction strength (P-value = 3.45e-06; one-sample Wilcoxon Test), suggesting that the differential changes in gene expression, PolII occupancy and interactions are independent of these TFs occupancy.

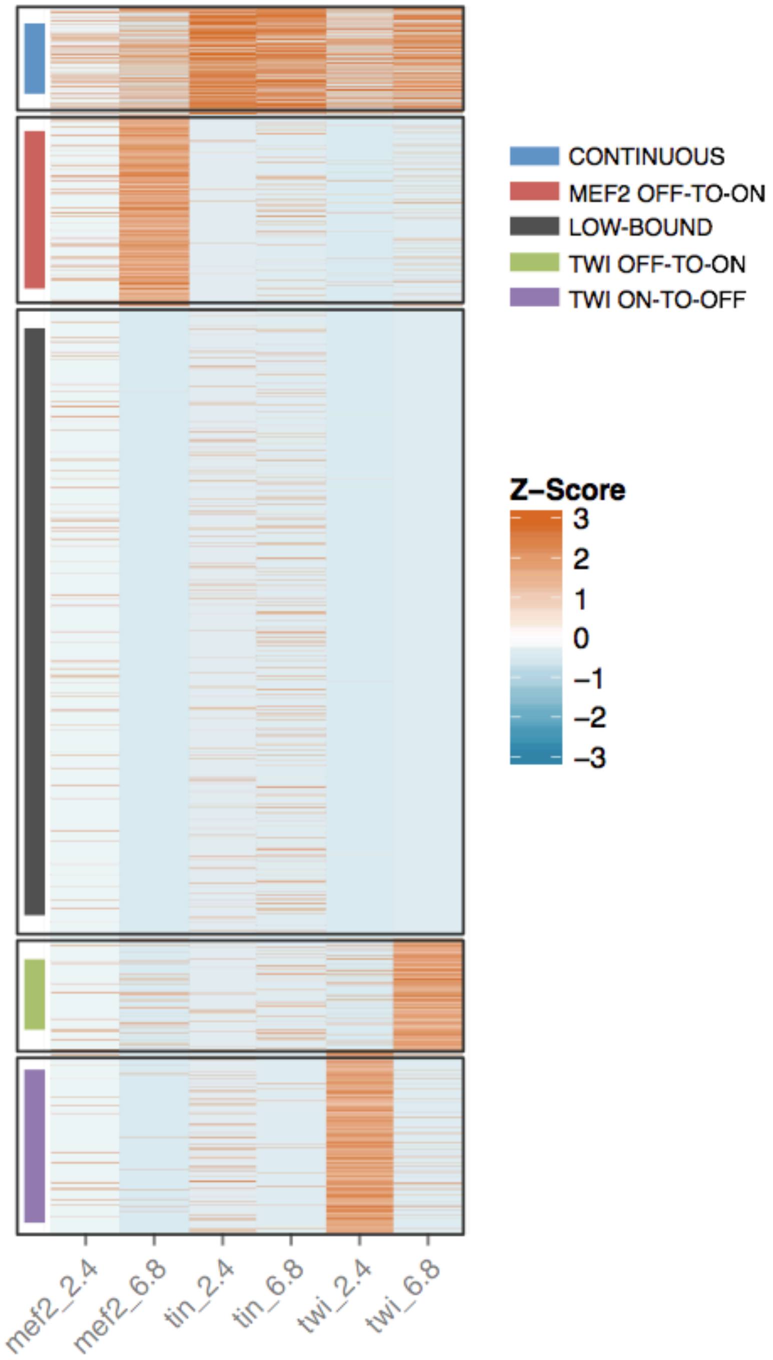

Figure 56. Occupancy of mesodermal factors on Hi-C enhancer anchors. Occupancy of three mesodermal transcription factors (Mef2, Tin and Twi), at two stages of embryonic development (3-4h, and 6-8h) on the Hi-C Enhancer anchors. K-means clustering of ChIP-chip enrichments converted to Z-scores reveals 5 distinct groups: continuous binding of all three factors (blue), temporal switch of Mef2 occupancy towards 6-8h stage (red), low-bound state (grey), and contrasting Twist transitions, where the occupancy is either increasing (green), or decreasing (purple) at the later developmental stage.

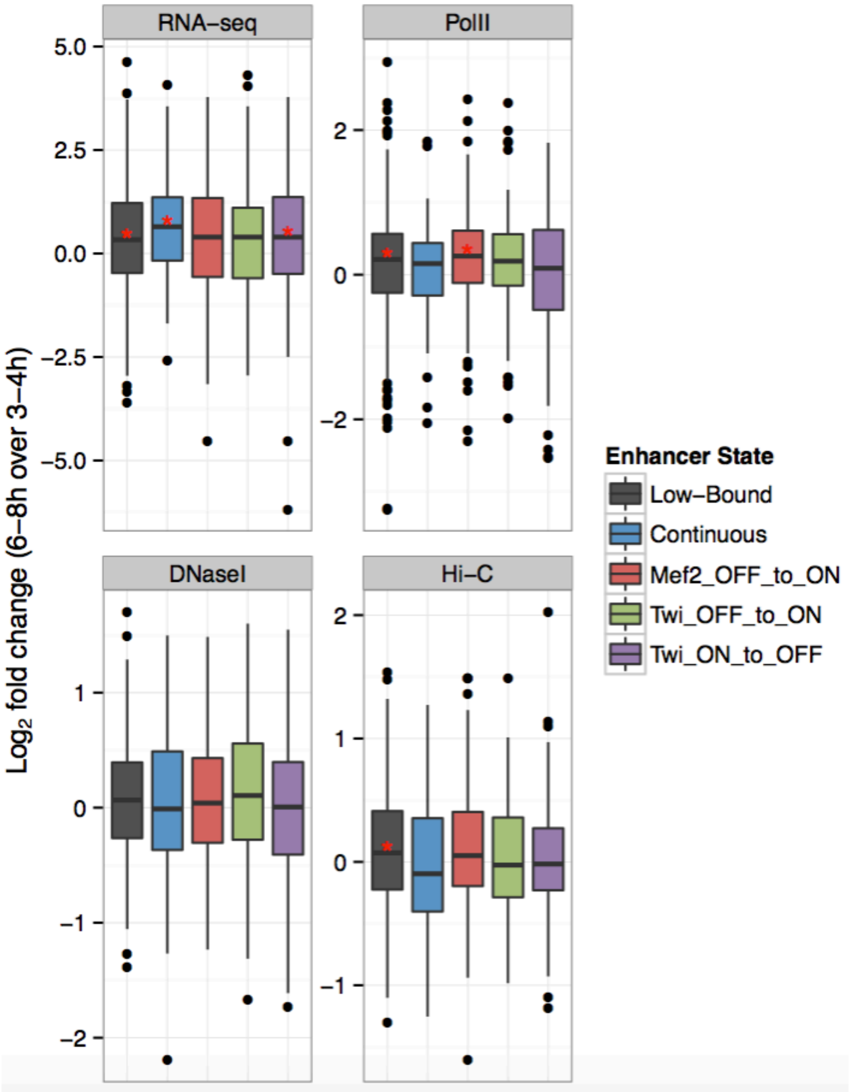

Figure 57. Dynamic correlations on interactions from varying enhancer anchor states. Differential correlations of enhancer anchor states defined by clustering at Figure 56, with four different datasets: gene expression (RNA-Seq), RNA Polymerase II occupancy (PolII), chromatin accessibility (DNaseI-hypersensitivity), and Hi-C differential contacts. Represented are the fold changes between later (6-8h), and earlier (3-4h) developmental stages. Significant deviations (having P-value lower than 0.001) from zero are shown with red stars above the median values (one-sample Wilcoxon Test).

4.3.5 Summary

In short, our Hi-C experiment shows high levels of reproducibility, quality and unprecedented resolution, which resulted in thousands of newly defined chromatin interactions emerging from enhancer and promoter defined anchors. This represents a significant increase in known contacts compared to the largest previous study, the 4C-study described in Chapter 2, providing the statistical power to assess many specific hypothesis involving the relationship between TF occupancy, chromatin state and three dimensional topology. In this chapter, I have focused on addressing the questions related to the global interactome (such as differences in interaction distance distribution between three anchor types: active promoters, non-active promoter and enhancer), functionality of the *Drosophila* interactions (including gene ontology analysis), epigenetic states and transcription factor occupancy, and differential interactions comparing both our experimental stages (3-4h and 6-8h), to the later embryonic stage (16-18h) from Cavalli lab. I have found that Hi-C interactions differ in their distance distributions, with ones originating from active promoters being shorter on average compared to enhancer and non-active anchors. They can also be found at very long range (above 100kb), although not as frequently as we previously observed using 4C, which is probably due to the technical sensitivity of this method.

On average, contacts tend to form hubs of similar features (for example, enhancer anchors frequently interact with other enhancers), an observation that is also supported by the shared epigenetic states between anchors and their interactions. Enhancer-based anchors show the greatest diversity in functionality of the genes they contact, recapitulating a lot of known categories for the measured embryonic stage such as morphogenesis, proliferation, growth and embryo development, compared to non-active promoter anchors that have almost no ontological enrichment. While there exists a low variation in number of significant Hi-C interactions across ontologically defined anchor categories, there is a higher tendency for promoters involved with locomotion (including cellular motility), and especially apoptosis (involved for example in formation of tissues during embryogenesis), to interact with developmental enhancers. Differential Hi-C analysis shows a higher differential potential when compared between the more distinct embryonic stages, while clustering of the transcription factor binding on enhancer anchors revealed 5 distinct groups of differing occupancy. Analysis of those 5 groups, involving the integration of external datasets (gene expression, chromatin accessibility, and RNA Polymerase

II occupancy), revealed that although some of the specific factors, such as Mef2, correlate with an increase in PolII occupancy, most interesting is the low-bound category, which shows an increase of interaction strength at the later time points, suggesting that other elements, such as different transcription factors (including insulators) and epigenetic marks might play a role in establishing the interactome.

5. DISCUSSION

5.1 Regulation and recruitment of Polycomb proteins on enhancers during early embryogenesis

Computational analyses of Polycomb repressive system recruiting proteins Pho and dSfmbt resulted in a range of novel findings related to the regulation of developmental enhancers during embryonic development. In particular, I have found that within the context of mesodermal development, Pho and dSfmbt occupy hundreds of unique genomic loci throughout the *Drosophila* genome. For both of these factors, binding loci are not uniformly and randomly spread throughout the genome, indicating a potential link between the occupancy and underlying regulation by Pho and dSfmbt. In particular, the global overview of whole-chromosome occupancy visualized by Hilbert curves revealed that both factors tend to cluster in particular regions that overlap with large-width domains of the repressive histone mark H3K27me3. Taking a more detailed look into data showed that PhoRC (constructed from the union of peaks of Pho and dSfmbt, which together form this protein complex) peaks tend to be preferentially associated with particular genomic features: promoter-proximal regions (within 1kb of gene TSS), and developmental enhancers (aggregated from various resources based on literature and TF occupancy; see Methods). The latter was especially intriguing, since most of the ChIP analyses up to date on Polycomb factors largely focused on promoters and PREs (Oktaba, K. et al. 2008; Kwong, C. et al. 2008; Schuettengruber, B. et al. 2009). Comprehensive analyses, including the careful selection of background regions, indicated that PhoRC may be functionally linked to the regulation of distal enhancer regions. This suggestion is supported by several lines of evidence: 1. Individual components of the PhoRC complex, Pho and dSfmbt, show higher levels of ChIP occupancy when in complex, rather than on individual sites; 2. Continuing on the previous result, when ChIP intensity is compared across different classes of genomic elements, developmental enhancers show significantly more reads, even when compared to promoters. Although the difference between actual protein affinity to the DNA sequence and amount of ChIP occupancy is still an open question, there is a reasonably strong correlation between the two, as demonstrated by the study on five regulators of early anterior-posterior patterning in *Drosophila* (~0.4 correlation; Kaplan, T. et al. 2011); 3. At least two-thirds of the enhancers that are occupied by PhoRC contain a *de novo* discovered Pho binding motif that resembles one previously reported in the literature (Oktaba, K. et al. 2008), indicating that the measured occupancy is likely due to direct interaction between DNA and Pho's zinc-finger binding domain; 4. Other proteins from the PRC1 complex (Pc and Ph) not only overlap with PhoRC on developmental enhancers, but also co-localise in higher frequency compared to promoter regions. Overall, the 225 PhoRC-bound developmental enhancers represent previously unappreciated recruitment loci for Polycomb-mediated repression. Since the repressive histone mark H3K27me3 that shows wide-spread signal when centered on PhoRC-bound developmental enhancers is catalytically deposited by the PRC2 protein E(z), it is quite likely that all three major groups of Polycomb proteins (PhoRC, PRC1 and PRC2) are present at distal regulatory regions; 5. Genes in the vicinity of PhoRC-occupied developmental enhancers show notably lower levels of gene expression compared to the reference set of active genes (e.g., ones nearest to the regulatory regions occupied by two or more enhancing mesodermal TFs). Albeit our ChIP experiments do not resolve the exact sequence of events that lead to repression of gene expression, results involving measures of TF occupancy, epigenetic marks and rate of transcription closely resemble the previously described and typical regulation by Polycomb system. Since our data is collected in mesoderm-specific cells, I have explored the option that Polycomb-based repression might directly influence the activity of mesodermal developmental enhancers. Indeed, I observe that PhoRC-bound regions show significantly lower levels of meso-sorted histone mark H3K27ac that is predictive of enhancer activity. This additional level of regulation might serve to further fine-tune the specification of tissues, especially within the later developmental stage when the overall morphology gets more complex. Differential analysis supports that role of Pho, since the highest dynamics was observed on enhancers, indicating that unlike in some regulatory contexts, such as Hox gene regulation, binding on enhancers might require more responsive and frequent exchange between repressive and activating TFs. Apart from the stereotypical repressive effect of Polycomb recruitment, these loci were also shown to be highly and significantly enriched for occupancy of the mesodermal transcription factor Twist. The importance of this co-occupancy was also confirmed through temporal profiling of time-specific Pho and dSfmbt signal on Twi-centred regions. Apart from the highest enrichment on the regions when Twi was bound at focused time point, an interesting pattern emerged indicative of priming the future occupancy. For the set of developmental enhancers that will be occupied by

Twi at 6-8h or later, but not at 4-6h or earlier in embryogenesis, both Pho and dSfmbt seem to have increased occupancy up to the level of Twi-bound regions, and noticeably higher than Pho 4-6h sample. This results does not support the hypothesis that Twi acts as a pioneer factor (this Twi functionality was suggested in Berkes, C. A. et al. 2004; Sandmann, T. et al. 2007) that would reveal binding site for Pho and thus correlate with Polycomb repression. Alternatively, due to the lack of any motif grammar between Pho and Twi, another possibility would be that Twi potentially stabilizes the PcG complex, or that the seemingly overlapping signal at 6-8h reflects the emerging complexity of mesodermal tissue that is in the process of being specified into cardiac, visceral and somatic muscle. My analyses of mammalian Pho homologue YY1 suggested that the discovered principle of recruitment on developmental enhancers might be evolutionary conserved from flies to humans: 1. Occupancy of YY1 peaks on DHS-based human developmental enhancers is both cell-specific, and statistically enriched compared to the matched background; 2. Some of those enhancers show epigenetic signatures that are characteristic for repressed regulatory regions and resemble the ones I observed in *Drosophila*, such as excess of H3K4me1 and H3K27me3, while lacking H3K4me3. To our knowledge, this is the first described occurrence of Polycomb recruitment and regulation of developmental enhancers in mesodermal embryonic development in *Drosophila*.

5.2 A search for candidate mammalian PREs, and hypothesis of a context-dependent regulatory switch between PREs and developmental enhancers

There is increasing evidence that PREs in mammals differ from *Drosophila*, whereby they lack a specific PRE TF signature, and are rather characterized by a combination of several loose properties, including CpG methylation, chromatin state and TF occupancy (Di Croce, L. & Helin, K. 2013; Riising, E. M. et al. 2014). Such a lack of clear PRE marks in mammals explains why only two examples of PREs has been identified to date (Woo, C. J. et al. 2010; Sing, A. et al. 2009; Cuddapah, S. et al. 2012). Our finding that YY1 significantly occupies enhancers in mammals raised two questions that we are currently addressing: 1. Could the dynamics of active and repressive histone marks on enhancers be used to identify candidate PREs?; 2. What are the differences in regulatory signatures that make the distinction between developmental enhancers and PREs? To address the first question, we are collaborating with Adrian Bracken from Trinity College Dublin, whose group has performed time-course ChIP experiments to describe the distributions of the H3K27ac active histone mark as well as of the Polycomb deposited H3K27me3 during mouse ES cell differentiation. Our hypothesis is that candidate PREs might be revealed by identifying changes from an activated to a repressed state of enhancers as differentiation progresses. Candidate regions would then be subjected to classical tests for their ability to function as Polycomb silencing elements, such as the ability of these regions to repress a mini-white reporter gene in transgenic assays in flies.

The second question, what is the distinction between enhancers and PREs, is fundamental to our understanding of regulatory sequences, and might explain the apparent discrepancy between the low number of tested PREs and our discovery that PcG protein occupancy is significantly enriched on developmental enhancers in both flies and humans. Our hypothesis is that a regulatory element can act both as a PRE and as a developmental enhancer, depending on the recruitment of specific TFs in a specific context. For example, such a region could function as an activator of target gene transcription if bound by a master regulator of mesodermal development Twist; or it could function as a repressor of target gene transcription if PcG proteins are recruited through PhoRC, as suggested by our results. To test that hypothesis, Raquel Marco-Ferreres in the Furlong lab, is experimentally testing if well characterised PREs can function as an enhancer in a transgenic enhancer-reporter assay. Conversely, some of the known developmental enhancers that I found to be occupied by both PhoRC components with H3K27me3 are being tested using two classical functional PRE tests: their ability to repress the function of an enhancer when placed in front of the *ubx* promoter, and their ability to silence a mini-white reporter gene. To summarise, our results suggest that PcG mediated repression can not only occur via promoter elements and ‘classic’ PREs, but also through a large number of developmental enhancers. This regulatory feature may be a conserved mode of action from *Drosophila* to humans, and help explain why there has been such limited PRE signatures identified in vertebrates to date.

5.3 Determination and functions of 4C interactions shaping the chromatin architecture of developmental enhancers

Many of the PhoRC regulated developmental enhancers are found far away from the closest TSS. Although nearest gene association is not a good predictor of target promoters (Sanyal, A. et al. 2012), it still means that those regulatory elements have to loop over large portions of non-coding genome to reach their targets, likely over other enhancers and genes in compact genomes such as *Drosophila*. Since the regulation by enhancers depends so strongly on the spatial topology within the nuclear space, I have performed analyses on chromatin interaction in the biggest 4C study up to date. Some of the significant interaction I have found were also part of chromatin looping from the Polycomb project. For example, viewpoint designed on eve promoter strongly interacts with distal eve enhancer eve_RP_eve_NR, which is also occupied by PhoRC, 8.5kb away from the eve promoter itself. But, apart from just looking at the repressive regulation, I integrated various other information on chromatin accessibility (DNaseI-Seq), gene expression (RNA-Seq), histone modifications (ChIP-Seq), transcription factor occupancy (ChIP-Seq) and others in order to understand the general principles underlying the topological organization of chromatin during embryonic development. These 4C interactions were found to be in concordance with the previously described enhancer-to-promoter contacts (for example, I recovered the known contact between *ap* promoter and ap_ApME680 enhancer). Surprisingly, I found that although *Drosophila* genome is quite compact, genomic interactions span very long genomic distances, often more than 50kb. This phenomenon is reminiscent of long-range enhancer regulation as observed in mammals, such as in the case of ZRS enhancer that is looping over 1Mb to reach the Shh gene it regulates (Lettice, L. A. et al. 2003). Long-range regulation in *Drosophila* also implies that many other enhancers and genes are skipped, suggesting that the genomic interaction are able to span regions independent of their underlying activity and function. Albeit enriched for active promoters and enhancers, similar to the previously reported findings in interactome studies (for example in Jin, F. et al. 2013), differential analyses between different spatiotemporal contexts that show considerable change in gene expression (~30%, estimated by the differential RNA-Seq analysis) revealed staggering stability of 4C interactions. Subset of stable contacts is occupied by high levels of paused RNA Polymerase II, as suggested by PolII-ChIP and GRO-Seq data. Since these interactions, which were termed as Differentially-Stable (DS), strongly correlate with the increase of target gene expression and increased transcription factor occupancy at the later point of embryogenesis, this topological organization might reflect a process of transcriptional bursting. For example, recently it was shown that RNA PolII is associated with the synchrony of temporal expression for major developmental genes such as even-skipped, and snail that controls the coordinated invagination of mesodermal cells during gastrulation (Lagha, M. et al. 2013; Bothma, J. P. et al. 2014). Our results support the topological model where enhancers and target promoters are placed in close spatial proximity well before the actual onset of transcription, and contain enriched levels of PolII, which can be triggered by the binding of specific transcription factors at the required temporal window.

5.4 Insights into topological organization of chromatin from high-resolution Hi-C data

Novel findings about the spatial organization of distal and promoter-proximal enhancer from the 4C study motivated me to further explore the topological organization of chromatin by analyzing all-to-all contact matrices from high-resolution Hi-C data. At the unprecedented resolution (with average interaction fragment being 200bp wide), I defined up to ~100,000 novel genomic interactions, with ~22,500 originating from annotated developmental enhancers and promoters. Distribution of anchor-to-interaction distances for various categorization of anchor regions (non-active promoter, active promoter and Enhancer) suggested shorter overall average distance length compared to the one observed from 4C-seq study (that is comparable due to the usage of same restriction enzyme DpnII). Careful probing of this result revealed that the observed discrepancy can probably be attributed to lack of sensitivity, rather than true biological difference. This conclusion is supported by several results: 1. Shorter 4C-seq interactions tend to have higher read counts, just as in the case of Hi-C data; 2. Albeit lower in read count frequency, 4C interactions longer than 100kb do not show a noticeable difference in statistical significance in comparison to the background model, which is in contrast to Hi-C where contacts with the distance of more than 100kb from the originating anchor have significantly higher P-value; 3. Long-range Hi-C interactions (above 100kb) have none or little supporting read counts. This result is overall not too surprising, since the number of anchors (4C 'viewpoints') in Hi-C data collection increased more than 3,500 fold (from 100 focused 4C-seq viewpoints to more than 350,000 Hi-C anchors). However, recovery of long-range Hi-C interactions could be possible with the implementation of further experimental and computational approaches: 1. Increased deep-sequencing and inclusion of filtered multi-fragment reads; 2. Aggregation of reads from neighboring anchors as the points of originating interactions. Despite the fact that it is challenging to detect long-range Hi-C interactions, remainder of the dataset still provided important and useful insights into the organization of *Drosophila* interactome. In particular, gene activity seems to play an important role in shaping the anchor-to-interactions distances. Active promoters have shortest contacts on average, followed by enhancers and significantly longer non-active promoter anchors. Hi-C interactions tend to be preferentially established between genomic elements of the same type, for example 14% of enhancer anchor interactions are other enhancers (compared to only 5% in active promoter, and 4% in non-active promoters). This result was additionally supported by the matching epigenetic signature between anchors and Hi-C contacts, and is consistent with the known observation that similar genomic elements (in terms of their regulatory function, and expression levels) tend to form hubs within the nuclear space, reminiscent of different nuclear structures such as Polycomb bodies and transcription factories (Sexton, T. et al. 2012; Pirrotta, V. & Li, H.-B. A 2012; Rieder, D. et al. 2012). Ontological analyses though clearly indicated that promoters that are in contact with developmental enhancers (enhancer anchors) are enriched in wider set of ontological categories (compared to anchors of other types), including biological processes such as multicellular organismal development, embryonic development, cell proliferation and morphogenesis. This result, despite the recent research showing striking similarity in architecture of promoters and enhancers including [production of transcripts, and binding of PolII, TATA-binding protein and TFIIB (Core, L. J. et al. (2014), suggests that non-coding regions might contain binding sites for transcription factors that in are in turn responsible for the control of genes with wider variety of biological functions. In line with these analyses that included external biological annotations, anchors with locomotion and apoptosis related functions seem to be particular interesting for several reasons: 1. Locomotion promoter anchors, which include genes responsible for cellular movement (migration) during the formation of tissue in development have both highest average interaction frequency and highest average anchor contacts; 2. Genes that could potentially trigger cell death have second-highest average frequency of enhancer interactions, albeit their overall contacts are not as frequent as in the case of locomotion. Both observation could reflect the need for multi-enhancer inputs that would provide both robust increase of gene expression in specific spatiotemporal contexts (locomotion), and the safeguard against potential mis-activation that would trigger cell death (apoptosis). Differential analysis between two developmental time points - 3-4h, and 6-8h of embryogenesis confirmed that very few Hi-C fragments change during the progression from multipotency to cell specification, as previously suggested by 4C-seq study, indicating that the later results were not biased toward a particular type of viewpoint regions. However, more differences are observed when our Hi-C data was compared to the one from Cavalli group, which was assessed at the later developmental time point (16-18h). Such outcome probably reflects the view that more changes in the topological architecture of the genome should be expected between more distant developmental time points, due to greater morphological and functional variability.

5.5 Conclusions and outlook

Overall, I have completed a variety of computational analyses that resulted in novel insights into the functional organization of regulatory elements within *Drosophila* genome. In particular, I have demonstrated that specific regulatory system, mediated by the transcription factors from Polycomb group proteins, is recruited to the developmental enhancers during early embryogenesis. This recruitment then lead to the repression of gene expression in the vicinity of developmental enhancers as demonstrated by overlap with the repressive histone mark H3K27me3 and transcriptional activity inferred from RNA-Seq data. Extending the approach to wider range of promoter-proximal and distal regulatory elements that were characterized using 4C-seq method, has shown that enhancers are remarkably stable across developmental progression, tend to be longer in genomic distance from target promoter than previously thought and poised for later activation by transcription factors. Advances in the sequencing technology, especially the inclusion of longer reads that would capture several interacting fragments, will provide stronger support for detecting interactions from Hi-C data. Together with the addition of datasets from the overlapping developmental stages and cell types, such integrative analyses would then enable a plethora of new opportunities in discovering novel functional links between the spatial organization of the genome and the underlying activity of genomic elements that together shape the developmental programmes in metazoan organisms.

6. REFERENCES

Aebi, U., Cohn, J., Buhle, L. & Gerace, L. The nuclear lamina is a meshwork of intermediate-type filaments. Nature 323, 560–564 (1986).

Aerts, S. Computational strategies for the genome-wide identification of cis-regulatory elements and transcriptional targets. Curr. Top. Dev. Biol. 98, 121–145 (2012).

Arnone, M. I. & Davidson, E. H. The hardwiring of development: organization and function of genomic regulatory systems. Development 124, 1851–1864 (1997).

Bailey, T. L. et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Research 37, W202–W208 (2009).

Bantignies, F. & Cavalli, G. Polycomb group proteins: repression in 3D. Trends in Genetics 27, 454–464 (2011).

Barozzi, I. et al. Coregulation of transcription factor binding and nucleosome occupancy through DNA features of mammalian enhancers. Molecular Cell 54, 844–857 (2014).

Baù, D. et al. The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. Nature Publishing Group 18, 107–114 (2011).

Beisel, C. & Paro, R. Silencing chromatin: comparing modes and mechanisms. Nat Rev Genet 12, 123–135 (2011).

Ben-Tabou de-Leon, S. & Davidson, E. H. Gene regulation: gene control network in development. Annu Rev Biophys Biomol Struct 36, 191 (2007).

Berkes, C. A. et al. Pbx marks genes for activation by MyoD indicating a role for a homeodomain protein in establishing myogenic potential. Molecular Cell 14, 465–477 (2004).

Biddie, S. C. et al. Transcription factor AP1 potentiates chromatin accessibility and glucocorticoid receptor binding. Molecular Cell 43, 145–155 (2011).

Bier, E. et al. Searching for pattern and mutation in the Drosophila genome with a P-lacZ vector. Genes Dev 3, 1273–1287 (1989).

Bird, A. DNA methylation patterns and epigenetic memory. Genes Dev 16, 6–21 (2002).

Bonn, S. et al. Cell type-specific chromatin immunoprecipitation from multicellular complex samples using BiTS-ChIP. Nature Protocols 7, 978–994 (2012).

Bonn, S. et al. Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. Nat Genet 44, 148–156 (2012).

Bothma, J. P. et al. Dynamic regulation of eve stripe 2 expression reveals transcriptional bursts in living Drosophila embryos. Proc Natl Acad Sci USA 111, 10598–10603 (2014).

Bracken, A. P. & Helin, K. Polycomb group proteins: navigators of lineage pathways led astray in cancer. Nat. Rev. Cancer 9, 773–784 (2009).

Brown, J. B. et al. Diversity and dynamics of the Drosophila transcriptome. Nature 1–7 (2014). doi:10.1038/nature12962

Brown, J. L., Fritsch, C., Mueller, J. & Kassis, J. A. The Drosophila pho-like gene encodes a YY1-related DNA binding protein that is redundant with pleiohomeotic in homeotic gene silencing. Development 130, 285–294 (2003).

Cedar, H. & Bergman, Y. Linking DNA methylation and histone modification: patterns and paradigms. Nat Rev Genet 10, 295–304 (2009).

Celniker, S. E. & Rubin, G. M. The Drosophila melanogaster genome. Annu. Rev. Genom. Human Genet. 4, 89–117 (2003).

Cerase, A. et al. Spatial separation of Xist RNA and polycomb proteins revealed by superresolution microscopy. Proc Natl Acad Sci USA 111, 2235–2240 (2014).

Chen, K. et al. A global change in RNA polymerase II pausing during the Drosophila midblastulatransition. eLife 2, (2013).

Chow, J. & Heard, E. X inactivation and the complexities of silencing a sex chromosome. Curr. Opin. Cell Biol. 21, 359–366 (2009).

Core, L. J. et al. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nat Genet 46, 1311–1320 (2014).

Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science 322, 1845–1848 (2008).

Courey, A. J. & Jia, S. Transcriptional repression: the long and the short of it. Genes Dev 15, 2786–2796 (2001).

Crawford, G. E. et al. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome Res 16, 123–131 (2006).

Cremer, T. & Cremer, M. Chromosome territories. Cold Spring Harb Perspect Biol 2, a003889 (2010).

Creyghton, M. P. et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci USA 107, 21931–21936 (2010).

Cuddapah, S. et al. A novel human polycomb binding site acts as a functional polycomb response element in Drosophila. PLoS ONE 7, e36365 (2012).

Daines, B. et al. The Drosophila melanogaster transcriptome by paired-end RNA sequencing. Genome Res 21, 315–324 (2011).

De Laat, W. & Duboule, D. Topology of mammalian developmental enhancers and their regulatory landscapes. Nature 502, 499–506 (2013).

de Napoles, M. et al. Polycomb group proteins Ring1A/B link ubiquitylation of histone H2A to heritable gene silencing and X inactivation. Developmental Cell 7, 663–676 (2004).

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

Degner, J. F. et al. DNase I sensitivity QTLs are a major determinant of human expression variation. Nature 482, 390–394 (2012).

Dekker, J., Marti-Renom, M. A. & Mirny, L. A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nature Publishing Group 14, 390–403 (2013).

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

Di Croce, L. & Helin, K. Transcriptional regulation by Polycomb group proteins. Nature Publishing Group 20, 1147–1155 (2013).

Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485, 376–380 (2012).

Dostie, J. et al. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res 16, 1299–1309 (2006).

Duan, Z. et al. A three-dimensional model of the yeast genome. Nature 465, 363–367 (2010).

Edelmann, P., Bornfleth, H., Zink, D., Cremer, T. & Cremer, C. Morphology and dynamics of chromosome territories in living cells. Biochim. Biophys. Acta 1551, M29–39 (2001).

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).

Engström, P. G., Ho Sui, S. J., Drivenes, O., Becker, T. S. & Lenhard, B. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. Genome Res 17, 1898–1908 (2007).

Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol 28, 817–825 (2010).

Essafi, A. et al. A wt1-controlled chromatin switching mechanism underpins tissue-specific wnt4 activation and repression. Developmental Cell 21, 559–574 (2011).

Falvo, J. V., Thanos, D. & Maniatis, T. Reversal of intrinsic DNA bends in the IFN beta gene enhancer by transcription factors and the architectural protein HMG I(Y). Cell 83, 1101–1111 (1995).

Fetting, J. L. et al. FOXD1 promotes nephron progenitor differentiation by repressing decorin in the embryonic kidney. Development 141, 17–27 (2014).

Filion, G. J. et al. Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. Cell 143, 212–224 (2010).

Finlan, L. E. et al. Recruitment to the nuclear periphery can alter expression of genes in human cells. PLoS Genet 4, e1000039 (2008).

Flicek, P. et al. Ensembl 2014. Nucleic Acids Research 42, D749–55 (2014).

Frankel, N. et al. Phenotypic robustness conferred by apparently redundant transcriptional enhancers. Nature 466, 490–493 (2010).

Fullwood, M. J. et al. An oestrogen-receptor-alpha-bound human chromatin interactome. Nature 462, 58–64 (2009).

Gaertner, B. et al. Poised RNA polymerase II changes over developmental time and prepares genes for future expression. Cell Rep 2, 1670–1683 (2012).

Gallo, S. M. et al. REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in Drosophila. Nucleic Acids Research 39, D118–D123 (2010).

Gavrilov, A. A., Golov, A. K. & Razin, S. V. Actual ligation frequencies in the chromosome conformation capture procedure. PLoS ONE 8, e60403 (2013).

Ghavi-Helm, Y. et al. Enhancer loops appear stable during development and are associated with paused polymerase. Nature 512, 96–100 (2014).

Gheldof, N. et al. Cell-type-specific long-range looping interactions identify distant regulatory elements of the CFTR gene. Nucleic Acids Research 38, 4325–4336 (2010).

Gibcus, J. H. & Dekker, J. The hierarchy of the 3D genome. Molecular Cell 49, 773–782 (2013).

Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R. & Lieb, J. D. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res 17, 877–885 (2007).

Gómez-Díaz, E. & Corces, V. G. Architectural proteins: regulators of 3D genome organization in cell fate. Trends Cell Biol. 24, 703–711 (2014).

Gorkin, D. U., Leung, D. & Ren, B. The 3D Genome in Transcriptional Regulation and Pluripotency. Stem Cell 14, 762–775 (2014).

Graveley, B. R. et al. The developmental transcriptome of Drosophila melanogaster. Nature 471, 473–479 (2011).

Guelen, L. et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature 453, 948–951 (2008).

Hakim, O. et al. Diverse gene reprogramming events occur in the same spatial clusters of distal regulatory elements. Genome Res 21, 697–706 (2011).

Handoko, L. et al. CTCF-mediated functional chromatin interactome in pluripotent cells. Nat Genet 43, 630–638 (2011).

He, H. H. et al. Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. Nat Meth 11, 73–78 (2014).

Hebenstreit, D. et al. RNA sequencing reveals two major classes of gene expression levels in metazoan cells. Mol. Syst. Biol. 7, 497 (2011).

Heger, P., Marin, B., Bartkuhn, M., Schierenberg, E. & Wiehe, T. The chromatin insulator CTCF and the emergence of metazoan diversity. Proc Natl Acad Sci USA 109, 17507–17512 (2012).

Ho, D., Imai, K., King, G. & Stuart, E. A. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. Journal of Statistical Software Vol. 42, 1–28 (2011).

Ho, J. W. K. et al. Comparative analysis of metazoan chromatin organization. Nature 512, 449–452 (2014).

Hong, J.-W., Hendrix, D. A. & Levine, M. S. Shadow enhancers as a source of evolutionary novelty. Science 321, 1314 (2008).

Hou, C., Li, L., Qin, Z. S. & Corces, V. G. Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. Molecular Cell 48, 471–484 (2012).

Hu, H. et al. CRL4B catalyzes H2AK119 monoubiquitination and coordinates with PRC2 to promote tumorigenesis. Cancer Cell 22, 781–795 (2012).

Hublitz, P., Albert, M. & Peters, A. H. F. M. Mechanisms of transcriptional repression by histone lysine methylation. Int. J. Dev. Biol. 53, 335–354 (2009).

Hübner, M. R., Eckersley-Maslin, M. A. & Spector, D. L. Chromatin organization and transcriptional regulation. Curr. Opin. Genet. Dev. 23, 89–95 (2013).

Hyde-DeRuyscher, R. P., Jennings, E. & Shenk, T. DNA binding sites for the transcriptional activator/repressor YY1. Nucleic Acids Research 23, 4457–4465 (1995).

Ji, H. et al. An integrated software system for analyzing ChIP-chip and ChIP-Seq data. Nat Biotechnol 26, 1293–1300 (2008).

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

Jin, F. et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature 503, 290–294 (2013).

Junion, G. et al. A transcription factor collective defines cardiac cell fate and reflects lineage history. Cell 148, 473–486 (2012).

Kaletta, T. & Hengartner, M. O. Finding function in novel targets: C. elegans as a model organism. Nat Rev Drug Discov 5, 387–398 (2006).

Kaplan, T. et al. Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early Drosophila development. PLoS Genet 7, e1001290 (2011).

Kasprzyk, A. BioMart: driving a paradigm change in biological data management. Database (Oxford) 2011, bar049 (2011).

Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics 26, 2204–2207 (2010).

Klingler, M., Soong, J., Butler, B. & Gergen, J. P. Disperse versus compact elements for the regulation of runt stripes in Drosophila. Developmental Biology 177, 73–84 (1996).

Kolasinska-Zwierz, P. et al. Differential chromatin marking of introns and expressed exons by H3K36me3. Nat Genet 41, 376–381 (2009).

Kulkarni, M. M. & Arnosti, D. N. Information display by transcriptional enhancers. Development 130, 6569–6575 (2003).

Kvon, E. Z. et al. Genome-scale functional characterization of Drosophila developmental enhancers in vivo. Nature 512, 91–95 (2014).

Kwong, C. et al. Stability and dynamics of polycomb target sites in Drosophila development. PLoS Genet 4, e1000178 (2008).

Lafarga, M. et al. Clastosome: a subtype of nuclear body enriched in 19S and 20S proteasomes, ubiquitin, and protein substrates of proteasome. Mol. Biol. Cell 13, 2771–2782 (2002).

Lagha, M. et al. Paused Pol II coordinates tissue morphogenesis in the Drosophila embryo. Cell 153, 976–987 (2013).

Lande-Diner, L. & Cedar, H. Silence of the genes--mechanisms of long-term repression. Nat Rev Genet 6, 648–654 (2005).

Lee, T. I. et al. Control of developmental regulators by Polycomb in human embryonic stem cells. Cell 125, 301–313 (2006).

Lenhard, B., Sandelin, A. & Carninci, P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nat Rev Genet 13, 233–245 (2012).

Lettice, L. A. et al. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Human Molecular Genetics 12, 1725–1735 (2003).

Lettice, L. A. et al. Opposing functions of the ETS factor family define Shh spatial expression in limb buds and underlie polydactyly. Developmental Cell 22, 459–467 (2012).

Lewis, E. B. A gene complex controlling segmentation in Drosophila. Nature 276, 565–570 (1978).

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

Li, H. Aligning sequence reads, clone sequences and assemblycontigs with BWA-MEM. arXiv 1–3 (2013).

Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).

Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326, 289–293 (2009).

Liu, J., Qian, L., Han, Z., Wu, X. & Bodmer, R. Spatial specificity of mesodermal even-skipped expression relies on multiple repressor sites. Developmental Biology 313, 876–886 (2008).

Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. Genome Biology 15, 550 (2014).

Lund, A. H. & Van Lohuizen, M. Polycomb complexes and silencing mechanisms. Curr. Opin. Cell Biol. 16, 239–246 (2004).

Malik, H. S. & Henikoff, S. Phylogenomics of the nucleosome. Nature Structural & Molecular Biology 10, 882–891 (2003).

Malmanche, N. & Clark, D. V. Drosophila melanogaster Prat, a purine de novo synthesis gene, has a pleiotropic maternal-effect phenotype. Genetics 168, 2011–2023 (2004).

Manning, L. et al. A resource for manipulating gene expression and analyzing cis-regulatory modules in the Drosophila CNS. Cell Rep 2, 1002–1013 (2012).

Maston, G. A., Landt, S. G., Snyder, M. & Green, M. R. Characterization of enhancer function from genome-wide analyses. Annu. Rev. Genom. Human Genet. 13, 29–57 (2012).

Mavrich, T. N. et al. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. Genome Res 18, 1073–1083 (2008).

Meyer, R. E. et al. Mps1 and Ipl1/Aurora B act sequentially to correctly orient chromosomes on the meiotic spindle of budding yeast. Science 339, 1071–1074 (2013).

Min, I. M. et al. Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. Genes Dev 25, 742–754 (2011).

Minoche, A. E., Dohm, J. C. & Himmelbauer, H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome Biology 12, R112 (2011).

modENCODE Consortium et al. Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science 330, 1787–1797 (2010).

Mohd-Sarip, A., Cléard, F., Mishra, R. K., Karch, F. & Verrijzer, C. P. Synergistic recognition of an epigenetic DNA element by Pleiohomeotic and a Polycomb core complex. Genes Dev 19, 1755–1760 (2005).

Mohn, F. et al. Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. Molecular Cell 30, 755–766 (2008).

Morey, L. & Helin, K. Polycomb group protein-mediated repression of transcription. Trends Biochem. Sci. 35, 323–332 (2010).

Morris, G. E. The Cajal body. Biochim. Biophys. Acta 1783, 2108–2115 (2008).

Müller, I., Boyle, S., Singer, R. H., Bickmore, W. A. & Chubb, J. R. Stable morphology, but dynamic internal reorganisation, of interphase human chromosomes in living cells. PLoS ONE 5, e11560 (2010).

Müller, J. & Kassis, J. A. Polycomb response elements and targeting of Polycomb group proteins in Drosophila. Curr. Opin. Genet. Dev. 16, 476–484 (2006).

Muramatsu, D., Singh, P. B., Kimura, H., Tachibana, M. & Shinkai, Y. Pericentric heterochromatin generated by HP1 protein interaction-defective histone methyltransferase Suv39h1. Journal of Biological Chemistry 288, 25285–25296 (2013).

Nagano, T. et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature 502, 59–64 (2013).

Nathan, D., Sterner, D. E. & Berger, S. L. Histone modifications: Now summoning sumoylation. Proc. Natl. Acad. Sci. U.S.A. 100, 13118–13120 (2003).

Naumova, N. et al. Organization of the mitotic chromosome. Science 342, 948–953 (2013).

Nègre, N. et al. A cis-regulatory map of the Drosophila genome. Nature 471, 527–531 (2011).

Nègre, N. et al. A Comprehensive Map of Insulator Elements for the Drosophila Genome. PLoS Genet 6, e1000814 (2010).

Noordermeer, D. et al. Temporal dynamics and developmental memory of 3D chromatin architecture at Hox gene loci. eLife 3, e02557 (2014).

Noordermeer, D. et al. The dynamic architecture of Hox gene clusters. Science 334, 222–225 (2011).

Nora, E. P. et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature 485, 381–385 (2012).

O'Sullivan, J. M. et al. Gene loops juxtapose promoters and terminators in yeast. Nat Genet 36, 1014–1018 (2004).

Ohler, U., Liao, G.-C., Niemann, H. & Rubin, G. M. Computational analysis of core promoters in the Drosophila genome. Genome Biology 3, RESEARCH0087 (2002).

Oktaba, K. et al. Dynamic Regulation by Polycomb Group Protein Complexes Controls Pattern Formation and the Cell Cycle in Drosophila. Developmental Cell 15, 877–889 (2008).

Ong, C.-T. & Corces, V. G. CTCF: an architectural protein bridginggenome topology and function. Nat Rev Genet 15, 234–246 (2014).

Ong, C.-T. & Corces, V. G. Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat Rev Genet 12, 283–293 (2011).

Panne, D., Maniatis, T. & Harrison, S. C. An atomic model of the interferon-beta enhanceosome. Cell 129, 1111–1123 (2007).

Park, P. J. ChIP-Seq: advantages and challenges of a maturing technology. Nat Rev Genet 10, 669–680 (2009).

Peng, Z., Mizianty, M. J., Xue, B., Kurgan, L. & Uversky, V. N. More than just tails: intrinsic disorder in histone proteins. Mol Biosyst 8, 1886–1901 (2012).

Pengelly, A. R., Copur, Ö., Jäckle, H., Herzig, A. & Müller, J. A histone mutant reproduces the phenotype caused by loss of histone-modifying factor Polycomb. Science 339, 698–699 (2013).

Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A. & Bejerano, G. Enhancers: five essential questions. Nature Publishing Group 14, 288–295 (2013).

Phillips-Cremins, J. E. & Corces, V. G. Chromatin Insulators:Linking Genome Organization to Cellular Function. Molecular Cell 50, 461–474 (2013).

Pietersen, A. M. & Van Lohuizen, M. Stem cell regulation by polycomb repressors: postponing commitment. Curr. Opin. Cell Biol. 20, 201–207 (2008).

Pirrotta, V. & Li, H.-B. A view of nuclear Polycomb bodies. Curr. Opin. Genet. Dev. 22, 101–109 (2012).

Pope, B. D. et al. Topologically associating domains are stable units of replication-timing regulation. Nature 515, 402–405 (2014).

Poux, S., McCabe, D. & Pirrotta, V. Recruitment of components of Polycomb Group chromatin complexes in Drosophila. Development 128, 75–85 (2001).

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

Rada-Iglesias, A. et al. A unique chromatin signature uncovers early developmental enhancers in humans. Nature 470, 279–283 (2011).

Rao, S. S. P. et al. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell 159, 1665–1680 (2014).

Reynolds, N., O'Shaughnessy, A. & Hendrich, B. Transcriptional repressors: multifaceted regulators of gene expression. Development 140, 505–512 (2013).

Rieder, D., Trajanoski, Z. & McNally, J. G. Transcription factories. Front Genet 3, 221 (2012).

Riising, E. M. et al. Gene silencing triggers polycomb repressive complex 2 recruitment to CpG islands genome wide. Molecular Cell 55, 347–360 (2014).

Ringrose, L. & Paro, R. Polycomb/Trithorax response elements and epigenetic memory of cell identity. Development 134, 223–232 (2007).

Risso, D., Schwartz, K., Sherlock, G. & Dudoit, S. GC-content normalization for RNA-Seq data. BMC Bioinformatics 12, 480 (2011).

Robinson, P. N., Wollstein, A., Böhme, U. & Beattie, B. Ontologizing gene-expression microarray data: characterizing clusters with Gene Ontology. Bioinformatics 20, 979–981 (2004).

Rosenbloom, K. R. et al. ENCODE data in the UCSC Genome Browser: year 5 update. Nucleic Acids Research 41, D56–63 (2013).

Rossetto, D., Avvakumov, N. & Côté, J. Histone phosphorylation: a chromatin modification involved in diverse nuclear events. Epigenetics 7, 1098–1108 (2012).

Rusconi, J. C., Hays, R. & Cagan, R. L. Programmed cell death and patterning in Drosophila. Cell Death Differ. 7, 1063–1070 (2000).

Sagai, T., Hosoya, M., Mizushina, Y., Tamura, M. & Shiroishi, T. Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. Development 132, 797–803 (2005).

Sandmann, T. et al. A core transcriptional network for early mesoderm development in Drosophila melanogaster. Genes Dev 21, 436–449 (2007).

Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. Nature 489, 109–113 (2012).

Saunders, A., Core, L. J., Sutcliffe, C., Lis, J. T. & Ashe, H. L. Extensive polymerase pausing during Drosophila axis patterning enables high-level and pliable transcription. Genes Dev 27, 1146–1158 (2013).

Scheuermann, J. C. et al. Histone H2A deubiquitinase activity of the Polycomb repressive complex PR-DUB. Nature 465, 243–247 (2010).

Schmitt, S., Prestel, M. & Paro, R. Intergenic transcription through a polycomb group response element counteracts silencing. Genes Dev 19, 697–708 (2005).

Schoeftner, S. et al. Recruitment of PRC1 function at the initiation of X inactivation independent of PRC2 and silencing. EMBO J. 25, 3110–3122 (2006).

Schuettengruber, B. et al. Functional Anatomy of Polycomb and Trithorax Chromatin Landscapes in Drosophila Embryos. Plos Biol 7, e13 (2009).

Schuettengruber, B., Chourrout, D., Vervoort, M., Leblanc, B. & Cavalli, G. Genome regulation by polycomb and trithorax proteins. Cell 128, 735–745 (2007).

Schwartz, Y. B. et al. Genome-wide analysis of Polycomb targets in Drosophila melanogaster. Nat Genet 38, 700–705 (2006).

Sexton, T. et al. Three-dimensional folding and functional organization principles of the Drosophila genome. Cell 148, 458–472 (2012).

Shiio, Y. & Eisenman, R. N. Histone sumoylation is associated with transcriptional repression. Proc. Natl. Acad. Sci. U.S.A. 100, 13225–13230 (2003).

Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: fromproperties to genome-wide predictions. Nat Rev Genet 15, 272–286 (2014).

Simon, J. A. & Kingston, R. E. Mechanisms of polycomb gene silencing: knowns and unknowns. Nat. Rev. Mol. Cell Biol. 10, 697–708 (2009).

Simon, J. A. & Lange, C. A. Roles of the EZH2 histone methyltransferase in cancer epigenetics. Mutat. Res. 647, 21–29 (2008).

Simonis, M. et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat Genet 38, 1348–1354 (2006).

Simonis, M., Kooren, J. & De Laat, W. An evaluation of 3C-based methods to capture DNA interactions. Nat Meth 4, 895–901 (2007).

Sing, A. et al. A vertebrate Polycomb response element governs segmentation of the posterior hindbrain. Cell 138, 885–897 (2009).

Smith, E. & Shilatifard, A. Enhancer biology and enhanceropathies. Nature Publishing Group 21, 210–219 (2014).

Sparmann, A. & Van Lohuizen, M. Polycomb silencers control cell fate, development and cancer. Nat. Rev. Cancer 6, 846–856 (2006).

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

Splinter, E. et al. CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. Genes Dev 20, 2349–2354 (2006).

Splinter, E. et al. The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. Genes Dev 25, 1371–1383 (2011).

St Pierre, S. E., Ponting, L., Stefancsik, R., McQuilton, P.FlyBase Consortium. FlyBase 102--advanced approaches to interrogating FlyBase. Nucleic Acids Research 42, D780–8 (2014).

Stock, J. K. et al. Ring1-mediated ubiquitination of H2A restrains poised RNA polymerase II at bivalent genes in mouse ES cells. Nat. Cell Biol. 9, 1428–1435 (2007).

Symmons, O. et al. Functional and topological characteristics of mammalian regulatory domains. Genome Res 24, 390–400 (2014).

Tan, M. et al. Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. Cell 146, 1016–1028 (2011).

Tavares, L. et al. RYBP-PRC1 complexes mediate H2A ubiquitylation at polycomb target sites independently of PRC2 and H3K27me3. Cell 148, 664–678 (2012).

Therizols, P. et al. Chromatin decondensation is sufficient to alter nuclear organization in embryonic stem cells. Science 346, 1238–1242 (2014).

Thomas, S. et al. Dynamic reprogramming of chromatin accessibility during Drosophila embryo development. Genome Biology 12, R43 (2011).

Thongjuea, S., Stadhouders, R., Grosveld, F. G., Soler, E. & Lenhard, B. r3Cseq: an R/Bioconductor package for the discovery of long-range genomic interactions from chromosome conformation capture and next-generation sequencing data. Nucleic Acids Research 41, e132–e132 (2013).

Tolhuis, B., Palstra, R.-J., Splinter, E., Grosveld, F. & De Laat, W. Looping and interaction between hypersensitive sites in the active beta-globin locus. Molecular Cell 10, 1453–1465 (2002).

Uslu, V. V. et al. Long-range enhancers regulating Myc expression are required for normal facial morphogenesis. Nat Genet 46, 753–758 (2014).

van Arensbergen, J., Van Steensel, B. & Bussemaker, H. J. In search of the determinants of enhancer-promoter interaction specificity. Trends Cell Biol. 24, 695–702 (2014).

van de Werken, H. J. G. et al. 4C technology: protocols and data analysis. Meth. Enzymol. 513, 89–112 (2012).

van de Werken, H. J. G. et al. Robust 4C-seq data analysis to screen for regulatory DNA interactions. Nat Meth 9, 969–972 (2012).

Villa, R. et al. Role of the polycomb repressive complex 2 in acute promyelocytic leukemia. Cancer Cell 11, 513–525 (2007).

Villar, D., Flicek, P. & Odom, D. T. Evolution of transcription factor binding in metazoans - mechanisms and functional implications. Nat Rev Genet 15, 221–233 (2014).

Vining, M. S., Bradley, P. L., Comeaux, C. A. & Andrew, D. J. Organ positioning in Drosophila requires complex tissue–tissue interactions. Developmental Biology 287, 19–34 (2005).

Visel, A., Rubin, E. M. & Pennacchio, L. A. Genomic views of distant-acting enhancers. Nature 461, 199–205 (2009).

Williams, R. L. et al. fourSig: a method for determining chromosomal interactions in 4C-Seq data. Nucleic Acids Research 42, e68–e68 (2014).

Williamson, I. et al. Spatial genome organization: contrasting views from chromosome conformation capture and fluorescence in situ hybridization. Genes Dev 28, 2778–2791 (2014).

Wittkopp, P. J. & Kalay, G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nat Rev Genet 13, 59–69 (2012).

Woo, C. J., Kharchenko, P. V., Daheron, L., Park, P. J. & Kingston, R. E. A region of the human HOXD cluster that confers polycomb-group responsiveness. Cell 140, 99–110 (2010).

Xiong, W., Dabbouseh, N. M. & Rebay, I. Interactions with the Abelson Tyrosine Kinase Reveal Compartmentalization of Eyes Absent Function between Nucleus and Cytoplasm. Developmental Cell 16, 271–279 (2009).

Yahata, T. et al. The MSG1 non-DNA-binding transactivator binds to the p300/CBP coactivators, enhancing their functional link to the Smad transcription factors. Journal of Biological Chemistry 275, 8825–8834 (2000).

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

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

Zhang, H., Levine, M. & Ashe, H. L. Brinker is a sequence-specific transcriptional repressor in the Drosophila embryo. Genes Dev 15, 261–266 (2001).

Zhang, W. I., Röhse, H., Rizzoli, S. O. & Opazo, F. Fluorescent in situ hybridization of synaptic proteins imaged with supe§r-resolution STED microscopy. Microsc. Res. Tech. 77, 517–527 (2014).

Zhou, V. W., Goren, A. & Bernstein, B. E. Charting histone modifications and the functional organization of mammalian genomes. Nat Rev Genet 12, 7–18 (2011).

Zhou, W. et al. Histone H2A monoubiquitination represses transcription by inhibiting RNA polymerase II transcriptional elongation. Molecular Cell 29, 69–80 (2008).

Zinzen, R. P., Girardot, C., Gagneur, J., Braun, M. & Furlong, E. E. M. Combinatorial binding predicts spatio-temporal cis-regulatory activity. Nature 461, 65–70 (2009).

7. APPENDIX

7.1 List of figures

7.2 List of tables

7.3 List of abbreviations

<	lower than
>	greater than
3C	Chromosome Conformation Capture
3D	three-dimensional
4C-seq	circular chromosome conformation capture followed by sequencing
5C	Carbon Copy Chromosome Conformation Capture
A	adenine
ac	acetylation
Bap	bagpipe
Bin	biniou
bp	basepair
BRG1	Brahma-related gene-1
C	cytosine
CAD	CRM Activity Database
ChIA-PET	Chromatin Interaction Analysis by Paired-End Tag Sequencing
ChIP-Seq	chromatin immunoprecipitation followed by sequencing
chr	chromosome
CpG	cytosine guanine base pairing (dinucleotide)
CRM	*cis* regulatory module
CTCF	CCCTC-binding factor
DHS	DNase I hypersensitive sites
DNA	deoxyribonucleic acid
DNaseI-Seq	DNase I hypersensitive sites sequencing
dSfmbt	Scm-related gene containing four MBT domains
EMBL	European Molecular Biology Laboratory
ENCODE	The Encyclopedia of DNA Elements
FACS	fluorescence-activated cell sorting
FAIRE-seq	Formaldehyde-Assisted Isolation of Regulatory Elements
FDR	false discovery rate
FISH	fluorescence in situ hybridization
G	guanine
GRN	gene regulatory network
GTF	general transcription factor
GWAS	genome-wide association study
H3	Histone 3
H4	Histone 4
kb	kilobase
log	logarithm
Mb	megabase
me	methylation
me2	di-methylation
me3	tri-methylation
Mef2	myocyte enhancer factor-2
mRNA	messenger ribonucleic acid
P300	E1A binding protein p300

Pc	polycomb
PcG	polycomb group of proteins
PCR	polymerase chain reaction
Ph	polyhomeotic
Pho	pleiohomeotic
PhoRC	pleiohomeotic repressive complex
PIC	pre-initiation complex
PolII	RNA Polymerase II
PRC1	polycomb repressive complex 1
PRC2	polycomb repressive complex 2
PRE	polycomb response element
RNA	ribonucleic acid
STARR-Seq	self-transcribing active regulatory region sequencing
T	thymine
TAD	topologically associated domain
TF	transcription factor
TF8008	enhancer list based on 8008 transcription factor sites
Tin	Tinman
TSS	transcription start site
Twi	Twist
Ub	ubiquitination
UCSC	University of California Santa Cruz
YY1	Yin Yang 1

7.4 Acknowledgments

Doctoral studies were an exceptionally challenging period of my life. The successful completion wouldn't be possible without the generous support from my advisors, colleagues, friends and family.

First of all, I would like to thank Eileen Furlong for all of the support, for giving me the opportunity to grow within the creative environment, and teaching me how to approach some of the most challenging questions in regulatory biology. In addition, I would like to thank members of my Thesis Advisory Committee - Dr. John Marioni, Prof. Dr. Michael Boutros and Dr. Wolfgang Huber for their valuable feedback.

Members of the Furlong group made my stay at EMBL Heidelberg particularly enjoyable. Especially, I cannot thank David Garfield enough for always being encouraging and approachable. I learned most valuable of life lessons from you. Most of my projects have been done in collaboration with other members within the group, including Jelena Erceg and Yad Ghavi-Helm, thank you. Also thanks to Charles Girardot, Matilda Maleš, Lucia Ciglar, and all other members I had chance to be in company of. It was a pleasure to work with all of you.

Lastly, I would like to thank my family and friends. Vi ste najzaslužniji za ovaj uspjeh. Zahvalio bih se prije svega svojoj obitelji - roditeljima Mirku i Barbari Pakozdi, te sestri Tanji Putar na njihovoj nesebičnoj podršci tijekom cijelog mojeg života. Ponosan sam na vas.

Posebno hvala dragom prijatelju Anti Lučiću. Puno smo toga zajedno prošli, ali najbolje nas još čeka. Zahvalan sam ti na svim trenutcima, smijehu i raspravama. Ovo je samo početak.