A genomic study of regulation of transcription by chromatin

Filipe Cadete
Fitzwilliam College
University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy
DECLARATION

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

No parts of this dissertation have been submitted for any other qualification.

This dissertation does not exceed the specified limit of 60,000 words as set by the Biology Degree Committee.

18 November 2013
Filipe Cadete
A genomic study of regulation of transcription by chromatin
Filipe Cadete
2013

All of an organism’s cells share the same genetic code. In the nucleus, the genome contains all the instructions necessary for life, from the most basic metabolic processes to intricate multi-tissue structures like the metazoan brain. Cell identity and proper development thus become a question of using the right set of instructions: expressing the correct set of genes in a timely fashion and at the right location. Gene regulation happens at several levels, including post-transcriptional stages, but the most important is arguably the control of transcription of DNA into RNA. DNA is not alone in the nucleus and interacts with a diverse set of proteins. This mixture of DNA and proteins is called chromatin and it is the immediate context in which transcription occurs. In this thesis I explore how transcription is regulated by chromatin.

*Drosophila* dosage compensation is a good model system for understanding how histone modifications fine-tune transcription across thousands of genes on an entire chromosome. Dosage compensation is the process by which differences in gene copy number are balanced in the final products of gene expression. In *Drosophila melanogaster* a twofold difference in the number of male and female sexual chromosomes is compensated by a twofold increase in the expression of genes on the male X. In collaboration with the Akhtar laboratory from the Max Planck Institute for Immunobiology and Genetics, I investigated the role of two components of the dosage compensation complex, Male Specific Lethal 1 (Msl1) and Males absent On the First (Mof), in two related *Drosophila* species. In doing so, I identified a novel genome-wide transcription-associated activity for Msl1. I also characterised the interaction of Dosage Compensation Complex (DCC) proteins with the RNA components of the complex, roX1 and roX2, using individual-nucleotide resolution cross linking and immunoprecipitation (iCLIP) datasets.

All the mechanisms of transcriptional regulation serve to influence the activity of the RNA polymerases. Most eukaryote genes are transcribed by RNA Polymerase II (Pol II). Using datasets generated by high-resolution chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-Seq) I identified five distinct modes of occupancy of Pol II in gene promoters in *Drosophila melanogaster*. I present evidence that these modes reflect the different transcriptional patterns of house-keeping and developmentally regulated genes. By further integrating datasets of transcription factor-binding, histone modifications, chromatin states or colours, nucleosome organisation and gene ontology annotations, I attempt to explain the factors underlying Pol II’s different patterns of occupancy.

The eukaryotic genome is highly packaged and organised in the nucleus. DNA makes long-range physical interactions in a non-random fashion that can connect gene promoters with distant enhancers or other regulatory regions, thus expanding the areas that potentially regulate a gene from proximal cis-elements to the entire genome. In recent years, a variety of techniques based on Chromatin Conformation Capture (3C) have allowed us to probe these interactions on a genome-wide scale. A new method developed by the Fraser laboratory from the Babraham Institute specifically queries interactions of all promoter regions with each other and with the rest of the genome. Using datasets produced by this new technique I investigated the chromatin landscape of regions interacting with promoters in both *Mus musculus* and *Homo sapiens*, in two different biological conditions for each species. I report a clear enrichment of transcription-related chromatin marks in regions interacting with highly-expressed promoters. These results will help elucidate the complex interplay between transcription regulation and three-dimensional genome organisation.

In this work I present several ways in which transcription is regulated by chromatin. Through greater understanding of these complex layers of gene expression, we will eventually fully describe and comprehend biological systems as commonplace as development or as harmful as cancer.
“No man is an island, entire of itself.” Since the start of my doctoral studies I have had the company of many, both close and far. I would like to acknowledge their help and support.

First of all, I want to acknowledge the European Molecular Biology Laboratory, the Max Planck Institute, Cancer Research UK and Epi-GeneSys for funding my work.

As an expatriate in a relationship with someone on a third country, knowing there is reliable transport to my loved ones kept me sane. Unbeknownst to them, TAP, BA and Eurostar have contributed to the completion of my studies.

My friends in Portugal have always made me feel like I haven’t left. I owe them apologies for the short notice I give them for my visits, my thanks for their friendship and a drink or two.

The collaborations with external laboratories put me in contact with fascinating colleagues. Among them, Asifa Akhtar, Sarantis Chlamydas, Ibrahim Ilik, Peter Fraser, Cameron Osborne and Stefan Schoenfelder especially have provided important counsel and debate in our work together.

An advantage of being part of a PhD programme is that one does not go through it alone. In the EMBL International PhD Programme I found fellow students that are now colleagues and friends. The predoc and friends community at the EBI, in particular, has made my time there enjoyable, both on and off work. I can only hope that they enjoyed reading my weekly lunch reminders as much as I enjoyed writing them. I will always remember my time in Cambridge and the friends I found there very fondly.

In my time in the Luscombe laboratory, it moved from Cambridge to London, new people joined, old people left. During it all, Nick Luscombe ensured it remained a pleasant work environment. I want to thank my labmates across the years for their help and company. And, of course, Nick for his guidance from the start.

As exciting as life abroad is, distance is a strain. Despite this, my family has always been there for me. I thank them for their love, support, patience and regular updates on the weather and cuisine from Portugal and Belgium.

Finally, I want to thank Claire for her love and companionship, both in person and on the other end of a Skype call. Maintaining a house is hard, we somehow managed to maintain two across the Channel.
# CONTENTS

List of Figures                                           xv
List of Tables                                            xvii
List of abbreviations                                     xix

1 INTRODUCTION                                           1

1.1 General description of eukaryotic transcription        1
1.1.1 Central dogma of molecular biology                  1
1.1.2 RNA polymerases carry out transcription            3
1.1.3 The DNA codes for the gene structure, promoters and various regulators 4
1.1.4 Steps for transcription initiation                  5

1.2 Transcription factors                                  6
1.2.1 Transcription factors directly bind DNA in a sequence-specific manner 6
1.2.2 TFs alter the environment around the promoters and influence the assembly of the transcription machinery 6

1.3 Histones                                              7
1.3.1 Histones form nucleosomes to package DNA           7
1.3.2 Nucleosome positioning plays a role in transcription 7
1.3.3 Post-translational modifications alter histones     8

1.4 Chromatin                                             9
1.4.1 The mixture of DNA, TFs, histones and other proteins forms chromatin 9
1.4.2 Chromatin structure helps regulate transcription   9
1.4.3 Chromatin is dynamic                               10
1.4.4 Chromatin is three-dimensional                     12

1.5 The DCC: a model for chromatin regulation             13
1.5.1 Dosage compensation in D. melanogaster is achieved by up-regulation of the male X chromosome 13
1.5.2 The DCC is responsible for dosage compensation    14
1.5.3 Acetylation of H4K16 is the hallmark of doubled transcription 15

1.6 Techniques used in today’s genomics                   16
1.6.1 Measuring the end result                           16
1.6.2 Mapping out protein-DNA interactions                18
1.6.3 Capturing conformation                             20
1.6.4 Making sense of data                               20

1.7 Statement of objectives                               21
1.7.1 The many lives of Msl1                              21

ix
1.7.2 lncRNAs in Drosophila dosage compensation ........................................... 21
1.7.3 Pol II distinct modes of action ................................................................. 21
1.7.4 Promoter capture leads to new insights in the genome 3D architecture ....... 22

2 THE OTHER LIFE OF THE drosophila dosage compensation complex ..... 23
2.1 Introduction ..................................................................................................... 23
  2.1.1 Dosage compensation in D. melanogaster is achieved by male X up-regulation . 23
  2.1.2 Mof is an acetyl-transferase responsible for H4K16Ac .................................. 23
  2.1.3 Msl1 directs the DCC to the X chromosome HAS ........................................ 24
  2.1.4 DCC members are present in other metazoans with very different dosage compensation mechanisms ................................................................. 24
  2.1.5 Objective .................................................................................................. 25
2.2 Results ............................................................................................................ 25
  2.2.1 Mof binding in D. melanogaster males ..................................................... 25
  2.2.2 Mof binding in D. virilis males ................................................................. 27
  2.2.3 Msl1 binding in D. melanogaster males .................................................... 29
  2.2.4 Msl1 binding in D. virilis males ................................................................. 30
  2.2.5 DCC binding is conserved between orthologs .......................................... 30
  2.2.6 Downsampling of D. virilis Mof dataset does not affect the balance between binding of X and autosomal genes .................................................. 33
  2.2.7 HAS sequence specificity is conserved .................................................... 34
  2.2.8 Msl1 binding in the absence of dosage compensation and Mof ............... 35
  2.2.9 Msl1 binding is associated with higher expression .................................... 38
  2.2.10 Msl1 depletion leads to reduction of Pol II Ser5 levels .............................. 40
  2.2.11 CDK7 inhibition alters Msl1 localization ................................................. 42
2.3 Discussion ..................................................................................................... 43

3 iCLIP REVEALS THE PROTEIN-RNA INTERACTION OF THE dosage compensation complex .................................................................................. 45
3.1 Introduction .................................................................................................. 45
  3.1.1 The D. melanogaster DCC and RNA ....................................................... 46
  3.1.2 iCLIP: probing protein-RNA interactions at the nucleotide level ............ 48
  3.1.3 Objective ................................................................................................. 48
3.2 Results ........................................................................................................... 49
  3.2.1 Mle binds DNA mainly in the X chromosome .......................................... 49
  3.2.2 Mle RNAi lowers the levels of all DCC components ................................. 52
  3.2.3 Mle binds mainly roX RNAs ..................................................................... 52
  3.2.4 Msl2’s main targets are also the roX RNAs ............................................. 54
3.2.5 roX1 and roX2 show defined regions bound by Mle and Msl2 ............... 55
3.2.6 Mle binds RNA stems ........................................... 57
3.2.7 SHAPE reveals roX structures ..................................... 57
3.2.8 roX2 stem structures are essential and redundant for male viability .............. 58
3.2.9 Each DCC only has one of the roX RNAs .................................. 60

3.3 Discussion ........................................................................ 62

4 RNA POLYMERASE II HAS DISTINCT MODES OF OCCUPANCY IN A HIGHER EUKARYOTE 65
4.1 Introduction ..................................................................... 65
4.1.1 Transcription as an interplay of multiple components ................................. 65
4.1.2 Drosophila melanogaster is a well-studied system ....................................... 65
4.1.3 An unbiased study from RNA Polymerase II's perspective ............................. 68
4.2 Results ............................................................................ 68
4.2.1 Analysis of Pol II datasets .................................................. 68
4.2.2 Pol II shows distinct profiles across TSSs ............................................... 69
4.2.3 Expression analysis .......................................................... 73
4.2.4 Clusters show different levels of transcriptional fine-tuning over the course of development ........................................... 74
4.2.5 Promoter types .............................................................. 75
4.2.6 Clusters show differential TF binding .............................................. 77
4.2.7 Nucleosome positioning .................................................... 79
4.2.8 Histone modifications and variants ................................................. 81
4.2.9 Chromatin types ............................................................ 83
4.2.10 GO analysis .................................................................... 85
4.2.11 Comparison with previous classifications .............................................. 86
4.3 Discussion ........................................................................ 88

5 PROMOTER CAPTURE-HI-C BRINGS NEW INSIGHTS TO CHROMOSOME CONFORMATION 91
5.1 Introduction ..................................................................... 91
5.1.1 3C-derived techniques capture chromosome conformation ................. 91
5.1.2 Different techniques for different questions ........................................... 92
5.1.3 The promoter capture-Hi-C technique ................................................. 93
5.2 Results ............................................................................ 93
5.2.1 General overview of the promoter C-Hi-C data ....................................... 93
5.2.2 Expression levels in the studied conditions ............................................. 96
5.2.3 Highly-expressed promoters interact with annotated enhancers ............... 96
5.2.4 Histone modifications in interacting regions reflect expression status .......... 97
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.2.5</td>
<td>Histone modification enrichment pattern is robust with regards to interaction distances</td>
</tr>
<tr>
<td>5.2.6</td>
<td>ChromHMM analysis</td>
</tr>
<tr>
<td>5.2.7</td>
<td>Cohabiting marks in interacting fragments follow known patterns</td>
</tr>
<tr>
<td>5.2.8</td>
<td>Promoter-promoter interactions</td>
</tr>
<tr>
<td>5.2.9</td>
<td>Topologically associated domains (TADs)</td>
</tr>
<tr>
<td>5.2.10</td>
<td>CTCF and cohesin</td>
</tr>
<tr>
<td>5.2.11</td>
<td>LADs are inert interaction-wise</td>
</tr>
<tr>
<td>5.3</td>
<td>Discussion</td>
</tr>
<tr>
<td>6</td>
<td>SUMMARY AND CONCLUSION</td>
</tr>
<tr>
<td>6.1</td>
<td>Summary</td>
</tr>
<tr>
<td>6.2</td>
<td>Conserved functions of Mof and Msl1 in <em>Drosophila</em></td>
</tr>
<tr>
<td>6.3</td>
<td>Protein-RNA interactions in fly dosage compensation</td>
</tr>
<tr>
<td>6.4</td>
<td>Distinct modes of Pol II binding in fly</td>
</tr>
<tr>
<td>6.5</td>
<td>C-HiC identifies promoter-interacting regions</td>
</tr>
<tr>
<td>6.6</td>
<td>Computational biology in today’s genomics</td>
</tr>
<tr>
<td>6.6.1</td>
<td>The importance of computational work in data analysis</td>
</tr>
<tr>
<td>6.6.2</td>
<td>Publicly available data and its impact</td>
</tr>
<tr>
<td>6.6.3</td>
<td>Wet-dry crosstalk</td>
</tr>
<tr>
<td>A</td>
<td>APPENDIX – COMPUTATIONAL METHODS</td>
</tr>
<tr>
<td>A.1</td>
<td>General software used</td>
</tr>
<tr>
<td>A.2</td>
<td>Methods for chapter 2</td>
</tr>
<tr>
<td>A.2.1</td>
<td><em>Drosophila</em> ChIP-seq processing</td>
</tr>
<tr>
<td>A.2.2</td>
<td>Assaying binding conservation</td>
</tr>
<tr>
<td>A.2.3</td>
<td>Expression analysis</td>
</tr>
<tr>
<td>A.3</td>
<td>Methods for chapter 3</td>
</tr>
<tr>
<td>A.3.1</td>
<td>MLE ChIP-seq</td>
</tr>
<tr>
<td>A.3.2</td>
<td>MLE and MSL2 iCLIP processing</td>
</tr>
<tr>
<td>A.4</td>
<td>Methods for chapter 4</td>
</tr>
<tr>
<td>A.4.1</td>
<td>Definition of Pol II clusters</td>
</tr>
<tr>
<td>A.4.2</td>
<td>Pol II clusters and expression status</td>
</tr>
<tr>
<td>A.4.3</td>
<td>Pol II clusters and core promoter elements</td>
</tr>
<tr>
<td>A.4.4</td>
<td>Pol II clusters and TF binding</td>
</tr>
<tr>
<td>A.4.5</td>
<td>Nucleosome analysis</td>
</tr>
<tr>
<td>A.5</td>
<td>Methods for chapter 5</td>
</tr>
<tr>
<td>A.5.1</td>
<td>Creation of Capture Hi-C interaction lists</td>
</tr>
<tr>
<td>A.5.2</td>
<td>Expression analysis</td>
</tr>
<tr>
<td>A.5.3</td>
<td>Interactions with enhancer-annotated fragments</td>
</tr>
<tr>
<td>A.5.4</td>
<td>Genomic interactions and chromatin modifications</td>
</tr>
<tr>
<td>A.5.5</td>
<td>Interactions in TADs</td>
</tr>
<tr>
<td>A.5.6</td>
<td>CTCF and cohesin analysis</td>
</tr>
</tbody>
</table>
A.5.7 Interactions and LADs . . . . . . . . . . . . . . . . . . . 159
| Figure 2.1 | Average gene profiles of Mof binding in male *D. melanogaster* larvae | 26 |
| Figure 2.2 | Average gene profiles of Mof binding in male *D. virilis* larvae | 28 |
| Figure 2.3 | Average gene profiles of MSL1 binding in male *D. melanogaster* larvae | 29 |
| Figure 2.4 | Average gene profiles of Msl1 binding in male *D. virilis* larvae | 31 |
| Figure 2.5 | Heat map of Mof and Msl1 binding in *D. melanogaster* and *D. virilis* orthologues | 32 |
| Figure 2.6 | Conservation of Mof binding between *D. melanogaster* and *D. virilis* | 33 |
| Figure 2.7 | Motifs of *D. melanogaster* and *D. virilis* HAS | 35 |
| Figure 2.8 | Average gene profiles of Msl1 binding in female *D. melanogaster* larvae | 36 |
| Figure 2.9 | Average gene profile of Msl1 binding in male *D. melanogaster* mutant larvae | 37 |
| Figure 2.10 | Expression of genes according to their Msl1 binding status | 39 |
| Figure 2.11 | Protein levels in Msl1 RNAi and control cells | 41 |
| Figure 2.12 | ChIP of Pol II Ser 5 and CDK7 in Msl1 targets | 42 |
| Figure 2.13 | Polytene chromosome stainings in CDK7 inhibited and control cells | 42 |
| Figure 3.14 | Mle binds DNA in the X chromosome | 50 |
| Figure 3.15 | Mle binds genes in the 3’ end | 51 |
| Figure 3.16 | DCC members are depleted upon Mle RNAi | 53 |
| Figure 3.17 | Mle binds mainly roX1 and roX2 transcripts | 54 |
| Figure 3.18 | Msl2 binds mainly roX1 and roX2 transcripts | 55 |
| Figure 3.19 | Mle and Msl2 bind specific domains of roX1 and roX2 | 56 |
| Figure 3.20 | Mle’s highest peaks are in double-stranded regions of roX2 | 57 |
| Figure 3.21 | SHAPE reveals the structure of roX RNAs | 59 |
| Figure 3.22 | Male viability is diminished when roX secondary structures are disrupted | 61 |
| Figure 3.23 | ChIRP suggests that only one type of roX is present in each complex in the nucleus | 62 |
| Figure 4.24 | Average profile of Pol II over all promoters for the male and female samples | 69 |
| Figure 4.25 | Pol II heat maps and profiles in salivary gland clusters | 70 |
Figure 4.26 Pol II clusters are robust

Figure 4.27 Pol II clusters in S2 are similar to those found in salivary glands

Figure 4.28 Cluster profile shape is independent of expression status

Figure 4.29 IQR of FPKM expression values for each cluster across developmental stages

Figure 4.30 Pol II clusters have different TF enrichments and binding locations

Figure 4.31 Nucleosome positioning and NFR size across all genes

Figure 4.32 +1 and -1 nucleosome focus

Figure 4.33 Nucleosome positioning across Pol II clusters

Figure 4.34 Pol II cluster enrichments in histone modifications

Figure 4.35 Chromatin colour of TSSs from the different Pol II clusters

Figure 5.36 Number and characteristics of promoter-other interactions

Figure 5.37 Expression increases with the number of promoter-enhancer interactions

Figure 5.38 Enrichments of interacting fragments overlapping with different modifications over all non-bait fragments

Figure 5.39 Enrichments of enhancer fragments overlapping with different modifications over all non-bait fragments

Figure 5.40 Distance control of enrichments of interacting fragments overlapping with different modifications for mouse samples

Figure 5.41 Distance control of enrichments of interacting fragments overlapping with different modifications for human samples

Figure 5.42 Distance control of enrichments of enhancer fragments overlapping with different modifications

Figure 5.43 Chromatin state enrichment in interacting fragments

Figure 5.44 Co-occurrence of histone modifications on interacting fragments

Figure 5.45 Chromatin modification on promoter-promoter interactions

Figure 5.46 TAD influence on interactions involving promoters

Figure 5.47 Insulator impact on promoter-other interactions

Figure 5.48 Most interactions of the mESC sample happen outside LADs
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 2</td>
<td>Number of genes bound by Mof in <em>D. melanogaster</em></td>
<td>27</td>
</tr>
<tr>
<td>Table 3</td>
<td>Number of genes bound by Mof in <em>D. virilis</em></td>
<td>28</td>
</tr>
<tr>
<td>Table 4</td>
<td>Number of genes bound by Msl1 in <em>D. melanogaster</em></td>
<td>30</td>
</tr>
<tr>
<td>Table 5</td>
<td>Number of genes bound by Mof in the downsampled <em>D. virilis</em> experiment</td>
<td>34</td>
</tr>
<tr>
<td>Table 6</td>
<td>Number of genes bound by Msl1 in <em>D. melanogaster</em> female larvae</td>
<td>36</td>
</tr>
<tr>
<td>Table 7</td>
<td>Number of genes bound by Msl1 in <em>D. melanogaster</em> mof2 mutant larvae</td>
<td>37</td>
</tr>
<tr>
<td>Table 8</td>
<td>Reads sequenced and mapped for ChIP-seq</td>
<td>49</td>
</tr>
<tr>
<td>Table 9</td>
<td>Number of TSSs in the salivary gland and S2 samples</td>
<td>72</td>
</tr>
<tr>
<td>Table 10</td>
<td>Number of expressed and stalled transcripts in each cluster and, in parentheses, the same value represented as the percentage of the transcripts of that cluster</td>
<td>74</td>
</tr>
<tr>
<td>Table 11</td>
<td>Proportion of TSSs with a predicted promoter type in its vicinity for each cluster</td>
<td>77</td>
</tr>
<tr>
<td>Table 12</td>
<td>Correspondence between Pol II clusters and gene patterns presented in (Ni <em>et al.</em>, 2010)</td>
<td>87</td>
</tr>
<tr>
<td>Table 13</td>
<td>Number of promoter-other interactions and baits involved</td>
<td>94</td>
</tr>
<tr>
<td>Table 14</td>
<td>Number of promoter-promoter interactions and baits involved</td>
<td>95</td>
</tr>
<tr>
<td>Table 15</td>
<td>Number of bait fragments in each expression class</td>
<td>96</td>
</tr>
<tr>
<td>Table 16</td>
<td>Number of fragments that interact with only one expression class</td>
<td>98</td>
</tr>
<tr>
<td>Table 17</td>
<td>Histone modification datasets used in mouse embryonic stem cells</td>
<td>114</td>
</tr>
<tr>
<td>Table 18</td>
<td>Histone modification datasets used in the analysis of mouse fetal liver cells</td>
<td>115</td>
</tr>
<tr>
<td>Table 19</td>
<td>Histone modification datasets used in the analysis of GM12878 cells</td>
<td>116</td>
</tr>
<tr>
<td>Table 20</td>
<td>Histone modification datasets used in the analysis of CD34+ cells</td>
<td>116</td>
</tr>
</tbody>
</table>
LIST OF ABBREVIATIONS

3C chromatin conformation capture
bp base pair
ChIP – Seq chromatin immunoprecipitation followed by high-throughput sequencing
C – Hi – C capture Hi-C
DCC dosage compensation complex
DNA deoxyribonucleic acid
iCLIP individual-nucleotide resolution cross linking and immunoprecipitation
HAS high-affinity site
mof males absent on the first
mle maleless
mRNA messenger RNA
msl1 male specific lethal 1
msl2 male specific lethal 2
nt nucleotide
PolII RNA Polymerase II
RNA ribonucleic acid
TAD topologically associated domain
TF transcription factor
TSS transcription start site
TTS transcription termination site
UTR untranslated region
roX RNA on the X
INTRODUCTION

1.1 GENERAL DESCRIPTION OF EUKARYOTIC TRANSCRIPTION

1.1.1 Central dogma of molecular biology

“It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.” This small sentence at the end of Watson and Crick’s 1953 Nature letter describing the molecular structure of deoxyribonucleic acid (DNA) was nothing but an understatement. In fact, the discovery of the double helix structure of DNA immediately suggested a mechanism by which the information encoded in it could be replicated. Since both strands of the helix are complementary, by separating them and recreating the complementary strand the information content of the original molecule could be passed onto two daughter molecules.

In a similar manner, the information coded in the DNA is passed on to another nucleic acid, ribonucleic acid (RNA). This is one of the steps of information transfer delineated in the central dogma of molecular biology (Crick, 1958, 1970). The “dogma” outlines the directions in which information is transferred between DNA, RNA and proteins. In simple terms, the information is passed from DNA to another DNA molecule (in a process called replication), from DNA to RNA (transcription) and from RNA to protein (translation).

There are less prevalent cases. For example, in retroviruses the information encoded in a RNA molecule can be reverse-transcribed into DNA and incorporated into a host’s genome. In this thesis I describe ways in which transcription is regulated by its nuclear environment. This first chapter will describe the process of transcription and how it is regulated in eukaryotic cells.

The genome contains the units of heritability

DNA contains all the information necessary for the proper functioning and development of all living organisms. The whole of an organism’s DNA is called the genome. A genome is physically separated into distinct DNA molecules, the chromosomes. In eukaryotes they reside in the nucleus. It is in this cellular compartment that DNA-related processes take place.

Even before DNA had been confirmed as the molecular basis for heritability of traits, there was an understanding that some form of discrete entities is responsible for this heritability. These discrete entities are called genes. Physically, rather than being individual molecules,
they are stretches of DNA that encode for RNA. The number of genes in an organism’s genome does not correlate perfectly with the organism’s complexity. The single-celled *Saccharomyces cerevisiae* has close to 7000 protein-coding genes, less than half of the roughly 20,700 present in the *Homo sapiens* genome, while the model plant *Arabidopsis thaliana* has around 27,000 genes.

*Gene expression needs to happen at the right time and place*

The genome contains all the information necessary for an organism’s development. In multicellular metazoa and in the absence of chimeras, all of an organism’s cells are descendants of a single zygote, and their genome is the result of the replication of that zygote’s genome. While somatic mutations and genomic rearrangements can occur during development, the genetic information in all the cells will be extremely similar. Yet cells differentiate and have different structures, functions and behaviours. This is achieved by regulating the expression of the genes. Gene expression is the process of transcribing and, if it is a protein-coding gene, translating the gene’s sequence to form its final product, RNA or protein.

Gene expression has to be temporally and spatially regulated. A gene whose product is involved in basic cellular functions, like those that encode for ribosomal components, needs to be expressed at all times and in all cells. These genes are so-called housekeeping genes. Other genes have a more restricted expression pattern. The genes that encode for the light-sensitive transmembrane proteins, for example, only need to be expressed in photoreceptor cells. The same reasoning applies to genes that trigger or control the development of organs or other multicellular structures: they need to be expressed in the right location and at the right time so that development is properly regulated. Otherwise, malformations and malfunctions can occur.

*Transcription as a regulatory step*

The transfer of information from DNA to the final gene product involves the steps of transcription and translation. Gene expression can therefore be regulated at many levels. For example, the rates of translation and of mRNA degradation will have an impact on the amount of protein produced. However, transcription is arguably the most important of the steps of gene expression, as it is the first one. For this reason, its regulation has been extensively studied.

Transcription itself is a complex process. It involves the coordinated recruitment of proteins to accessible areas of the genome. Each recruitment step or each event that makes the required DNA sequences more or less accessible is a possible point of regulation. Here I describe the regulation of transcription in broad terms with a special focus on regulation via the chromatin environment.
1.1.2 RNA polymerases carry out transcription

The RNA polymerases are at the core of the transcription machinery. These are large protein complexes with several subunits that carry out the synthesis of RNA from a DNA template. There are different types of RNA polymerases in eukaryotic cells (Roeder and Rutter, 1969), with a structurally conserved core of 10 subunits between them (Vannini and Cramer, 2012). The three RNA polymerases transcribe different kinds of genes. RNA Polymerase I (Pol I) transcribes only ribosomal RNAs and RNA Polymerase III (Pol III) transcribes small ribosomal 5S RNA, tRNA and other small RNAs. Protein coding genes are transcribed by RNA Polymerase II (Pol II), and it is the regulation of its activity that this thesis will focus on.

RNA Pol II is composed of 12 subunits, RPB1 to RBP12 (Myer and Young, 1998). The two largest subunits, RPB1 and RPB2, are located opposite each other, with a cleft between them where DNA is inserted and where the synthesis of the new RNA molecule occurs. The other, smaller, subunits are located around the the structure formed by RPB1 and RPB2 (Cramer et al., 2000). RPB1 has a C-terminal domain (CTD) that contains a repeat sequence of seven amino-acid residues that is necessary for the proper functioning of Pol II (Brickey and Greenleaf, 1995). Specific residues in the CTD repeats can be phosphorylated and these post-translation modifications play an important role in the regulation of Pol II activity during the transcription cycle. The phosphorylation of serine 2 and of serine 5 in particular plays important roles in transcription elongation and initiation, respectively.

During transcription, Pol II takes on several roles. This protein complex maintains the DNA at its core in a single-stranded state. As the polymerase moves through the gene, it catalyses the addition of a new nucleotide to the nascent RNA molecule that matches the DNA template. In addition to its RNA synthesis function, the structure of Pol II allows it to backtrack when it detects a mismatch between the DNA template and the recently added RNA nucleotide (Sydow et al., 2009) and to subsequently resume transcription (Cheung and Cramer, 2011), giving it proofreading capability.

In addition to the subunits of Pol II, a set of general transcription factors (GTFs) participate in transcription. First identified in 1980, GTFs are required for proper initiation of transcription by Pol II (Matsui et al., 1980). They are known as TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIHH. They can themselves be multi-protein complexes. GTFs help Pol II assemble at promoters and coordinate its release into the gene to start transcription.
1.1.3 The DNA codes for the gene structure, promoters and various regulators

Both genes and their regulatory regions are encoded in the DNA sequence. At the 5' end of the gene we find a region called the promoter. It is in this region that the RNA polymerase machinery assembles and proceeds to transcribe the gene in a 5' to 3' direction. Given that it is where the first step of transcription happens, the promoter contains several regulatory elements to control the start of gene expression.

Most eukaryotic genes contain introns and exons. Introns are spliced out from the transcribed RNA and do not contain coding information. Exons contain the codons necessary for translation and are preserved in the processed messenger RNA (mRNA). The intron-exon border contain specific nucleotide sequences that mark them as such and are recognised by the splicing machinery. In most genes, the initial and final regions of the processed mRNA do not encode for amino acids and are known as untranslated regions (UTRs). The end of the gene, after the 3' UTR, varies between the genes that are translated by the different polymerases. In most protein-coding genes, a specific sequence contains the signal for the formation of a long adenine tail, the poly(A) tail.

In addition to the classical gene structure, there are other regulatory elements that are not necessarily located in genic regions. Some of these elements are enhancers that regulate gene expression, which, as we will see later, can act across long linear distances, possibly by DNA loops that bring the promoter and the enhancer close in the three-dimensional space of the nucleus. Enhancers are characterised by certain proteins and histone modifications.

Basic promoter sequence architecture

The promoter region is where the Pol II holoenzyme assembles to start transcribing a gene. There are core sequence motifs present at metazoan promoters that are recognised by the basic transcription machinery and help define the transcription start site (TSS).

Some of these elements appear in all metazoans. Such elements include the initiator (Inr), the TATA box and the upstream and downstream B-recognition elements (BREu and BREd, respectively). Inr is located inside the TSS where it is recognised by TFII B ([Kostrewa et al., 2009]). The TATA box is, as the name suggests, enriched for thymine and adenine. It lies upstream of the TSS and is recognised by the TATA box binding protein (TBP) that is part of TFII D. Like the Inr, BRE elements are recognised by TFII B. BREu and BREd differ in their relative locations, with the former being upstream and the latter downstream of the TATA box.

Other core promoter elements are not common to all metazoans ([Lenhard et al., 2012]). The three downstream core elements (DCE1,
DCE2 and DCE3) are present in vertebrates, and appear, as the name implies, downstream of the TSS. In fly there are also other downstream elements, the motif ten element (MTE) and the downstream promoter element (DPE). Fly promoters can also contain the DNA recognition element (DRE).

Not all promoters of the same organism contain all the core promoter elements. In fly, for example, the core elements have been grouped in five modules that contain one or two different core elements (Ohler, 2006). In a similar manner, vertebrate promoters can be classified according to the presence of a TATA box and to the GC content of their sequence (Carninci et al., 2006). The different types of promoters are important for transcription regulation. Genes that are expressed throughout development and those that are expressed only at specific times or tissues are enriched for different classes of promoters.

1.1.4 Steps for transcription initiation

Assembly of the transcription machinery

Pol II depends on the GTFs for assembly at the promoter. In a TATA box promoter, this element is bound by the TFIIID subunit TBP. In TATA-less promoters, it is another complex, SAGA, that recruits TBP. TBP recruitment leads to the recruitment of TFIIA and TFIIIB (Buratowski et al., 1989). These two GTFs stabilise the interaction of TBP with DNA. Meanwhile, TFIIF interacts with Pol II in the nucleus. Pol II is directed to the promoter via contacts that it and TFIIF establish with TFIIIB. After Pol II recruitment, TFII E recruits TFIIH. The assembled Pol II and GTFs at the promoter form the pre-initiation complex (PIC).

Release of Pol II

Once the PIC is assembled, a series of steps take place to start transcription. TFIIH plays a crucial role in this process. Its helicase activity separates the two strands of DNA to create a region of single-stranded DNA at the location of Pol II binding that is called a transcription bubble (Wang et al., 1992). TFIIH also contains the CDK7 kinase, which phosphorylates serine 5 of the CTD (Komarnitsky et al., 2000). The phosphorylation of serine 5 leads to the initiation of transcription and to the recruitment of the proteins that will add a guanine nucleotide to the 5’ end of the nascent mRNA, a process called capping. Capping of mRNA increases its stability inside the cell and prevents its degradation. During elongation, the CTD also becomes phosphorylated on serine 2 (Komarnitsky et al., 2000), which will lead to the recruitment of a histone methylase associated with transcription elongation.
**Pol II stalling after initiation is widespread**

Shortly after transcription initiation Pol II can arrest its progress through the gene and stop, still bound to the DNA, a phenomenon called stalling. This behaviour is widespread throughout the genome (Muse et al., 2007). As with its recruitment to the promoter, Pol II stalling is a regulated process. Stalling is enhanced by complexes such as the negative elongation factor (NELF), which in turn has its effect repressed by the positive transcription elongation factor b (P-TEFb, Peterlin and Price, 2006). It also seems to be more prevalent at developmental control genes which have specific expression patterns, suggesting that pausing plays a role in the regulation of their transcription (Zeitlinger et al., 2007).

### 1.2 TRANSCRIPTION FACTORS

#### 1.2.1 Transcription factors directly bind DNA in a sequence-specific manner

The GTFs that recruit Pol II to promoter don’t have specificity for DNA sequences beyond the core promoter elements. A different class of transcription factors (TFs) is sequence-specific. These contain DNA-binding domains (DBD), such as helix-turn-helix or zinc finger domains, which are responsible for their sequence-specificity and for the binding to DNA (Luscombe et al., 2000). TFs also contain effector domains that will carry out their regulatory functions.

The sequences recognised by the DBDs of TFs can vary in size and are usually between 6 and 12 bp long. The presence of a TF binding site is not by itself a guarantee that a TF is binding there. The context of the binding site plays a role in whether it has a regulatory function, depending on how close it is to the promoter, the accessibility of the DNA and the nature of the TF that binds to it.

There are around 1,400 TFs in the human genome (Vaquerrizas et al., 2009), less than one tenth of the total number of genes. In order to regulate those genes they have to act together, both in a combinatorial fashion — two or more TFs binding to and regulating the expression of the same promoter — and in a hierarchical fashion — TFs regulating the expression of other TFs. This way, they form complex transcriptional regulatory networks (Babu et al., 2004).

#### 1.2.2 TFs alter the environment around the promoters and influence the assembly of the transcription machinery

The effector domains carry out the TFs’ regulatory functions. Effector domains can act in different ways to impact gene expression, depending on the nature of their binding partners. When interacting with
the basal transcription machinery, the effector domains can activate
gene expression by helping recruit the general transcription factors
to promoters. Alternatively, a TF can interact with other sequence-
specific TFs. This can have the effect of stabilising the interaction
between the TFs and DNA. Finally, effector domains can help recruit
chromatin-modifying enzymes. These modify the chemical nature and
the positioning of chromatin-structuring proteins which, as we will see
later, impacts DNA accessibility and transcription (Frietze and Farn-
ham, 2011).

The proteins that a TF recruits to promoters, be they other TFs or
chromatin-modifying enzymes, can also inhibit transcription. In that
case, the TF acts as a repressor. The existence of both activator and
repressor TFs increases the complexity of gene regulatory networks, as
a single promoter can be the target of several TFs competing among
themselves.

1.3 histones

1.3.1 Histones form nucleosomes to package DNA

In eukaryotes DNA interacts with a specific type of proteins to form
higher-order structures. These proteins are called histones and they
assemble into octamers, called nucleosomes (Kornberg, 1977). There
are four distinct types of histones in each nucleosome (H2A, H2B, H3
and H4) with each nucleosome containing two of each type. Around
each of these octamers are wrapped 145-147 bp of DNA, in 1.65 turns
around the nucleosome (Luger et al., 1997). In addition to the core
global domains of the histones that form the structure around which
DNA wraps itself, histones also have flexible N-terminal tails that pro-
trude out of the nucleosome core. In a similar way to Pol II’s CTD,
the histone tails can be post-translationally modified to regulate gene
expression.

1.3.2 Nucleosome positioning plays a role in transcription

Nucleosome occupancy at promoters occludes or reveals binding sites
and regulates access to DNA

Nucleosome density is not identical across the genome. Their posi-
tioning is a regulated process that impacts and reflects DNA-related
activities. Around promoters, they take well-defined positions relative
to the TSS. In yeast, a nucleosome is part of the TSS (Albert et al.,
2007). In higher metazoans such as fly, the TSS does not overlap with
a nucleosome, but there is one just downstream (Mavrich et al., 2008).
By convention, the nucleosome at or just downstream of the TSS is
the +1 nucleosome and the nucleosome upstream of it is the -1 nucle-
osome. The stretch of DNA between the +1 and the -1 nucleosomes is known as the nucleosome-free region (NFR). As the location of the +1 nucleosome changes between species, so does the size of the NFR. In yeast this is roughly 150 base pairs long, while in flies it is wider, due to the downstream positioning of the +1 nucleosome. It is in this region that the core promoter elements and the promoter-proximal TF binding sites are usually located. By moving the -1 nucleosome up- or downstream, the NFR can vary in size and thus allow the binding sites to be bound or not.

\textit{Nucleosomes can include non-standard histones}

While most nucleosomes are made of the normal four histone types, other histone variants exist. These variants can replace the normal histones in the octamer. Nucleosomes that contain the special histone variants are found at specific locations of the genome and are associated with DNA-related processes.

Of particular significance to transcription is the H2A.Z variant of histone H2A. This variant is present at the +1 and -1 nucleosomes of yeast \cite{Albert2007} and at the +1 nucleosome of fly promoters \cite{Mavrich2008} and is associated with actively transcribed or poised promoters. Variant H3.3 of histone H3 is also linked to actively transcribed regions, as it replaces histone H3 after it is evicted by a RNA polymerase. The combination of H2A.Z and H3.3 is not restricted solely to promoters, and is also present at enhancers \cite{Jin2009}.

\subsection*{1.3.3 Post-translational modifications alter histones}

The N-terminal tails of histones can be post-translationally modified. These modifications usually consist of the addition of a small chemical group or of a peptide to a amino-acid residue of the histone tail \cite{Kouzarides2007}.

The chemical groups usually added in histone modifications are methyl groups (CH$_3$, in a process known as methylation), acetyl groups (COCH$_3$, acetylation), phosphate groups (PO$_4^{3-}$, phosphorylation). Of these, methylation is a particular case, as a specific amino-acid residue can be modified with between one and three methyl groups. As for peptides, the two most commonly added to histone tails are ubiquitin (ubiquitylation) and SUMO (sumoylation). Not all amino-acid residues are susceptible to post-translational modification. Lysines are the most versatile: they can be acetylated, methylated, ubiquitinated and sumoylated. Others are more restricted. Serines, for example, are usually phosphorylated.

Post-translation modification of histone tails is not a passive process. Each modification is catalysed by a specific class of enzymes. For example, acetylation of amino-acid residues is done by acetyltransferases, and the removal of an acetyl group is done by a deacetylase. The
1.4 Chromatin

1.4.1 The mixture of DNA, TFs, histones and other proteins forms chromatin

The components previously described do not act by themselves. Transcription and its regulation arises from the interactions between sequence-specific TFs, GTFs, the RNA polymerases, the DNA sequence and the nucleosomes, with all their variations and modifications. This mixture of DNA and proteins that are present in the cell nucleus is called chromatin.

The basic unit of chromatin is the nucleosome. As DNA wraps around successive histone octamers, it resembles the much-evoked image of “beads on a string”. This “beaded string” can coil and bend to form higher-order structures. Some of these structures will be highly wound and compacted, while others will be looser.

1.4.2 Chromatin structure helps regulate transcription

The compactness of chromatin has a direct effect on the context in which processes such as transcription occur. Using chromatin-staining imaging, two distinct forms have been observed. Heterochromatin shows higher staining and is heavily compacted. It is associated with
gene-poor, lowly-expressed regions and sits on the periphery of the nucleus. Euchromatin, on the other hand, stains less in microscopy experiments and represents a looser form of compaction. Euchromatic regions are gene-rich, with a large proportion of expressed genes.

Heterochromatin and euchromatin are not necessarily absolute states and chromatin can change between them. Whole chromosomes can change from one state of chromatin to the other: female mammals inactivate one of their two copies of the X-chromosome by inducing its compaction. The polycomb repressive complex, involved in the regulation of the developmental Hox genes, among others, induces the formation of heterochromatin through the methylation of the lysine 27 on the histone H3 to repress its target genes. A different, constitutive form of heterochromatin is formed by heterochromatin protein 1 (HP1) which binds to trimethylated histone 3 lysine 9 (H3K9me3).

1.4.3 Chromatin is dynamic

Composition changes from region to region

In addition to the broad categories of euchromatin and heterochromatin, chromatin can be divided into a number of states depending on its particular protein content. As mentioned above, heterochromatin can be divided into areas marked with H3K27me3 and those marked with H3K9me3. Euchromatic regions can also differ among themselves in the proteins that bind them and the modifications they carry, with functional implications.

As we have seen, promoters are the area where the transcription machinery assembles and, as such, can participate in the regulation of transcription. In addition to the binding of TFs and GTFs, active promoters carry particular histones. The aforementioned H2A.Z and H3.3 are two of them. Another is the trimethylation of lysine 4 in the histone H3 (H3K4me3, Santos-Rosa et al., 2002; Pokholok et al., 2005). Downstream of promoters, the methylation status of H3K4 changes from trimethylation to dimethylation and, further downstream, to monomethylation. Promoters are also marked by the acetylation of lysine 9 in the histone H3 (H3K9ac, Pokholok et al., 2005).

Other non-promoter regulatory regions in euchromatin have a particular protein composition. Enhancers are long-distance regulators that are bound by the P300 protein, an acetyl transferase also involved in acetylating promoters, and contain high levels of H3K4me1 and H3K27ac (Visel et al., 2009; Heintzman et al., 2007; Creyghton et al., 2010).

Promoter protein makeup changes during transcription initiation

The variability in the protein composition is not only observed between different areas of the genome. The same area undergoes changes in its
protein makeup during transcription. Promoters are a prime example of how change happens.

We have already seen how the binding of GTFs occurs in a step-wise manner that leads to the assembly of Pol II and to its subsequent phosphorylation and initiation of transcription, and how there is a NFR at promoters. This status is not permanent and depends on the activity of other protein complexes and on the presence and absence of particular marks.

To establish the NFR that makes the promoter sequence accessible any nucleosomes present there need be moved or evicted. Four non-sequence-specific complexes actively remodel the chromatin by changing the position of nucleosomes. They have to be actively recruited into the promoter area, possibly by sequence-specific TFs. Chromatin remodelers can also be recruited through post-translational modification of enzymes. The SWI/SNF family of chromatin remodelers contains bromodomains, which bind acetylated lysines (Hassan et al., 2001, 2002). As promoters contain nucleosomes with H3K9ac, this may help direct the remodelers to the promoter area. Three of these complexes are associated with active transcription: SWI/SNF, INO80/SWR1 and CHD (Bao and Shen, 2007). Not all chromatin remodelling complexes create a NFR to enable transcription. The fourth family of remodelers, ISWI, arranges nucleosomes in a manner that leads to a repression of transcription (Goldmark et al., 2000; Fazzio et al., 2001).

The assembly of the PIC is accompanied by the sequential modification of promoter nucleosomes. The phosphorylation of serine 5 of the CTD triggers the recruitment of proteins that eventually ubiquitinate H2BK123 in yeast and H2BK120 in vertebrates and Drosophila. This in turn leads to the recruitment of SET1 that catalyses H3K4 trimethylation (Dover et al., 2002; Sun and Allis, 2002). Downstream of the TSS, the change of H3K4me3 to H3K4me2 (Pokholok et al., 2005) is accompanied by the change in the phosphorylation status of the CTD. This phosphorylation of CTD’s serine 2 (Komarnitsky et al., 2000) also leads to the recruitment of SET2 and the methylation of H3K36 (Li et al., 2002; Kizer et al., 2005). These are not the only cases of interaction between the basal transcription machinery, the chromatin remodelers and histone modifications, but they illustrate how this crosstalk is a dynamic element that leads to changes of protein composition of promoters during transcription.

**DNA itself can be methylated**

It is not only proteins that can be chemically modified. The nucleotide residues cytosine and adenine can have a methyl group added to them. As with histone modifications, this process does not occur randomly. DNA methylation changes over the course of an organism’s development and will influence the activity of the methylated regions. Specifically, highly-methylated regions of the genome are not expressed. Si-
lencing via DNA methylation is used to maintain gene repression after cell division.

The level and extent of DNA methylation is not the same in different organisms. The Drosophila DNA, for example, is not methylated and methylation of adenosines only occurs in prokaryotes. In mammals most of the genome’s cytosines are methylated, and the areas that contain a low number of methylated cytosines are associated with regulatory elements (Stadler et al., 2011).

1.4.4 Chromatin is three-dimensional

Chromosomes form their own domains

Chromatin can vary in more than its protein content. The long fibres that make up each chromosome are tightly packaged inside the nucleus. In so doing, regions of the genome that might be distant from each other in the DNA sequence, or even on different chromosomes altogether, can be in close proximity in the three-dimensional environment of the nucleus.

The first order of three-dimensional organisation of the nucleus are the chromosomes themselves. They are clearly distinguishable as separate entities during mitosis. During interphase, while they do not form such tight condensed structures, chromosomes still occupy distinct territories inside the nucleus (Bolzer et al., 2005).

Inside chromosomes, there are topological domains

Each chromosome is divided into topological domains. These are discrete megabase-sized regions along the genome where most interactions are between loci inside the same region. Topological domains have been observed in yeast (Duan et al., 2010), fly (Sexton et al., 2012), mouse and human (Dixon et al., 2012).

Despite the presence of domains, long-range interactions still occur that connect loci in different domains. Interestingly, loci in domains that have a higher degree of transcriptional activity tend to interact preferentially with loci in other such active domains (Sexton et al., 2012).

Topological domains are conserved across cell types and species. Their borders are enriched for insulator proteins such as CTCF and for housekeeping genes. Boundaries between two topological domains can mark an abrupt separation between a heterochromatic region and euchromatin (Dixon et al., 2012). Current knowledge suggests that topological domains are a form of high-order chromatin organisation just below the chromosome territories.
Long-range interactions add an extra layer to transcription regulation

In addition to the overarching organisation of chromosomal and topological domains, small loci also form long-distance interactions that add another layer of complexity to the nuclear environment. Two types of interactions in particular have an impact in transcription: those between two promoters and those between a promoter and an enhancer.

The number of nascent transcripts and active RNA polymerases far outnumbers the number of discrete foci where transcription is occurring in the nucleus (Jackson et al., 1998), suggesting that each of these foci contains a high number of polymerases. These concentrations of transcription machinery are called transcription factories and are thought to help coordinate the expression of jointly-regulated genes. When a gene is transcribed, it loops out of its domain to contact a transcription factory (Osborne et al., 2004). By bringing promoters physically closer, transcription factories can help regulators interact with several promoters in a more efficient manner. In pluripotent stem cells, for example, the binding sites of pluripotency TFs Oct4 and Nanog form clusters in the three-dimensional nucleus (de Wit et al., 2013).

In addition to contacts between promoters, contacts between a promoter and an enhancer also play a role in regulating transcription. Enhancer activity is cell-type specific and plays a role in the activation of genes required for cell fate decision (Heintzman et al., 2009). Enhancers can be located away from the promoter and form loops to contact their targets (Mishiro et al., 2009). This interaction between an enhancer and its promoter target is created by the protein complexes cohesin and mediator (Kagey et al., 2010).

The interactions between promoters and between a promoter and an enhancer add an extra layer of regulation to gene expression. Promoters tend to interact more with areas of the genome that are open or that contain histone modifications that mark for transcriptional activity (Sanyal et al., 2012). The full impact of the three-dimensional organisation of the nucleus at a genomic level is not yet fully understood, but the emergence of new experimental and computational techniques will soon allow for that full understanding.

1.5 THE DOSAGE COMPENSATION COMPLEX: A MODEL FOR CHROMATIN REGULATION

1.5.1 Dosage compensation in D. melanogaster is achieved by up-regulation of the male X chromosome

Metazoan species with dimorphic sex chromosomes face an imbalance in their gene dosage, as one gender will have two identical copies of one of the sexual chromosomes while the other gender will have one of each of the forms of the sex chromosome. In flies and vertebrates females have two copies of the X chromosome and males have one copy of the
X and one copy of Y, a small, gene-poor chromosome. If each copy of a gene that is present in the X but not in the Y chromosome produces the same amount of final gene product, then females will have twice the amount of gene product than males. Dosage compensation is the process by which this imbalance is corrected.

The mechanism of dosage compensation is different across metazoans. Early during vertebrate development one of the female X chromosomes is condensed into repressive heterochromatin by the action of HP1, Polycomb group complexes and DNA methylation in a process initiated by the expression of non-coding RNA X-(inactive)-specific transcript (XIST). Once one of the female X chromosomes is fully repressed, males and females alike have only one active X chromosome and dosage compensation is achieved. In fly dosage compensation relies on an increase of transcription, not on repression. The expression from the male X chromosome is up-regulated two-fold, so that the same amount of gene product is obtained from one copy of a gene in the male X and from two copies of the same gene from the female X chromosomes (Straub and Becker, 2007).

In *D. melanogaster* the X chromosome contains over 15% of all protein-coding genes. The up-regulation of such a sizeable proportion of the genome, together with the ease of use of this model organism, makes fly dosage compensation a good system for studying large-scale transcription regulation.

In addition to primary tissue, cell lines have also been used in the study of gender-specific behaviour in *D. melanogaster*. S2 cells have been derived from the late embryonic stages (Schneider, 1972) and, due to their male phenotype of DCC formation and activity, they have been used in studies aiming to characterise the dosage compensation mechanism in *Drosophila* (for example, in Akhtar et al. (2000); Zhang et al. (2010); Gelbart et al. (2009)).

### 1.5.2 The DCC is responsible for dosage compensation

The dosage compensation complex (DCC) responsible for establishing the increase in transcription in the male X chromosomes of flies is composed of a core of five proteins and two non-coding RNAs. The protein members of this complex are male-specific lethal 1, 2 and 3 (Msl1, Msl2 and Msl3), maleless (Mle) and males absent on the first (Mof). The two non-coding RNAs are RNA on the X 1 and 2 (roX1 and roX2).

DCC members are expressed in both sexes, but the complex itself forms only in males. Sex-lethal (xxl), which is not part of the DCC, has differential splicing in male and female flies. In its female form, Sxl inhibits the translation of *msl2* mRNA, preventing the formation of the DCC. In males, the translated Msl2 protein interacts directly with Msl1, stabilising it (Lyman et al., 1997; Copps et al., 1998). Msl1 then
recruits Mof and Msl3 into the complex (Morales et al., 2004; Kadlec et al., 2011). Msl2 also activates the transcription of roX1 and roX2 (Rattner and Meller, 2004), which are then added to the DCC by Mle (Meller et al., 2000).

The assembled DCC targets sites on the X chromosome and spreads along the DNA from these sites in cis, until it covers the entire chromosome (Kelley et al., 1999). The chromatin entry sites, or high-affinity sites (HAS), contain a GA-rich motif that is enriched in the X chromosome compared to autosomes, providing sequence-specificity for DCC targeting to the X (Straub et al., 2008; Alekseyenko et al., 2008).

### 1.5.3 Acetylation of H4K16 is the hallmark of doubled transcription

The DCC contains an acetyl-transferase: Mof. This enzyme specifically acetylates lysine 16 of histone H4 (H4K16ac). This mark is not restricted to the male X chromosome of fly. It is present at promoters of autosomal genes in fly (Kind et al., 2008; Gelbart et al., 2009), and in mice it is found at active genes and enhancers (Taylor et al., 2013). However, this acetylation is most extensive on the male X chromosome of fly (Kind et al., 2008; Gelbart et al., 2009). H4K16ac inhibits the formation of compact chromatin fibres in vitro (Shogren-Knaak et al., 2006) and is associated with areas of higher DNA accessibility in vivo (Bell et al., 2010).

The broad acetylation of H4K16 nucleosomes of the male X chromosome suggests that the observed two-fold increase in transcription is achieved via more accessible chromatin. Two hypotheses of how Pol II is affected by H4K16ac have been proposed. In one of the models, Pol II is loaded equally at promoter of X and autosomal genes. The more accessible chromatin allows for better elongation of the mRNA and it is the higher processivity of Pol II in acetylated genes that leads to increased transcription on the X chromosome. Measuring the production of nascent mRNA in S2 cells (Larschan et al., 2007) observed this increase in Pol II elongation and no increase in Pol II promoter loading between X and autosomes.

The other hypotheses is that Pol II has the same rate of elongation on X and autosomal promoters, but is recruited at a higher rate to begin with on X-linked promoters. This was shown to be the case using ChIP-seq of Pol II in male and female fly salivary glands. In females, the amount of Pol II present along the gene — including the promoter — is identical between X-linked and autosomal genes. By contrast, males show an increase of Pol II in X-linked genes along the entire gene, promoter included. No difference was observed in the elongation of X-linked and autosomal genes (Conrad and Akhtar, 2012; Vaquerizas et al., 2013). Measuring nascent transcripts only as in Larschan et al. (2007) captures active polymerases. ChIP-seq is therefore more suited to compare the presence of Pol II in the promoter and the gene body.
Dosage compensation in *D. melanogaster* exemplifies not only transcription regulation on a genomic level, but also a way in which this regulation is achieved through the modification of the chromatin content and structure. Thus it is an invaluable model system for the study of widespread transcriptional regulation via chromatin modifications.

1.6 Techniques used in today’s genomics

1.6.1 Measuring the end result

The last decade saw an increase in the number of available genomes and sequencing capability. With these new tools, biologists have increasingly looked at the functional behaviour of the genome as a whole, compared to the previous approach of using smaller, model loci. The new techniques that have been developed probe different aspects of nuclear activity. To understand transcription, an essential measure is the amount of end product (RNA) produced.

Expression profiling by array

Before the advent of massively parallel sequencing there was already a way to measure the expression levels of several genes at once. Using a reverse transcriptase, mRNA can be converted into complementary DNA (cDNA). DNA probes can be created that complement the cDNA of the genes that are to be tested for transcription. These probes are printed onto a surface to which the cDNA is added in conditions that facilitate the hybridisation between the probes and the cDNA. By measuring the amount of cDNA that hybridised to each probe the original amount of mRNA in the cell can be inferred.

The first experiments based on this technique in the late 1980s used small arrays printed onto filter paper (Kulesh *et al.*, 1987). Expression profiling as a modern genome-wide tool became available when technological advances led to the miniaturisation of the array itself (Schena *et al.*, 1995) and when the sequencing of the genome of whole organisms allowed for the creation of probes for most of its genes (Lashkari *et al.*, 1997). Microarrays used for expression profiling usually contain probes mapping only to coding regions of the genome. Arrays that cover the genome more extensively (tiling arrays) are also available for experiments where the extent of the genome that is assayed extends further than coding regions.

The raw readout of a microarray experiment is the fluorescence of each of its probes. After quality control, a series of computational techniques are used to convert this signal into a measure of gene expression. The first step of data processing is background correction to eliminate the noise from image acquisition. If there are replicates or several arrays to compare, there is a normalisation step between the arrays to make the values comparable to each other (Gautier *et al.*, 2004). Some tools
identify what genes are differentially expressed between experimental conditions (Smyth, 2004). Others simply assign presence of absence calls to individual probes (Hubbell et al., 2002). Using this technique one can measure and compare expression levels of known genes, with the restriction that only the genetic area spotted on the array can be assayed.

RNA-Seq

As DNA sequencing technology became faster and cheaper, it was applied to purposes beyond genome sequencing. RNA-seq is a technique in which the amount of RNA present in a cell is measured by massively parallel sequencing. The RNA extracted from the cells is not sequenced directly; as in microarray expression profiling it is reverse transcribed into cDNA before sequencing.

The precise protocol for RNA-seq varies depending on the nature of the RNA that is to be sequenced. If the objective is to identify processed mRNA, a step for selection of poly(A) tails is included (Mortazavi et al., 2008; Nagalakshmi et al., 2008). Other protocols create their sequencing libraries from small RNAs to identify non-coding RNAs like micro RNAs. Compared with microarray expression profiling, RNA-seq has a larger dynamic range and improved sensitivity at the extremes of this range (Wang et al., 2009). Its lack of hybridisation probes also allows it to detect novel exons, intron/exon junctions and transcripts.

As a technique based on massively parallel sequencing, the output of the experiment is a collection of short DNA sequences (usually 25 to 200 bp), with an accompanying score reflecting the confidence of the calling of each nucleotide. Similar to microarray expression profiling, computational processing is needed to convert the raw output into expression levels. Quality control takes into account the sequencing confidence scores and the distribution of the four bases in the sequenced sample. The sequenced reads are then mapped onto the reference genome using specialised mapping software that takes into account that sequenced reads can include long gaps when mapped to the reference genome if they span an exon/intron boundary (e.g., Trapnell et al., 2009). From the mapped reads the levels of expression of a given transcript can be inferred, usually by measuring the number of reads that map to a transcript per kilobase of transcript per million mapped reads (RPKM). The number of reads that map to each transcript can also be compared across samples to identify differentially expressed transcripts (e.g., DESeq, Anders and Huber, 2010) or individual exons (e.g., DEXSeq, Anders et al., 2012).

The versatility of RNA-seq and its independence of probes make it an important tool for the study of the transcriptome. However, its processing methods are not yet as standardised as those of microarray expression profiling. If a researcher’s aims include the discovery of new
transcripts or an investigation into differential genome usage, RNA-seq is the tool of choice. If expression levels are all that’s required, a microarray experiment will in most cases be a simpler and cheaper alternative.

1.6.2 Mapping out protein-DNA interactions

Determining the location of proteins that bind DNA or otherwise interact with chromatin is essential to understand how they help regulate transcription. Several techniques have been developed that identify protein binding sites.

ChIP-Seq probes proteins and DNA

Chromatin Immunoprecipitation followed by sequencing (ChIP-seq) is a technique to determine protein-DNA binding sites genome-wide. It relies on the cross-linking of DNA and proteins to establish strong chemical connections between the two. This cross-linking is usually achieved with a chemical agent such as formaldehyde. The cross-link products are then subjected to sonication to shear the DNA into small fragments. This pool of small DNA fragments cross-linked to proteins is subjected to immunoprecipitation using an antibody that targets the protein of interest. As a result of the immunoprecipitation the mixture is enriched for DNA fragments that are cross-linked to the target protein. The cross-link is then reversed and the resulting DNA fragments are sequenced. Before massively parallel sequencing became widespread, immunoprecipitation was followed by hybridisation to a tiling array (ChIP-chip). Usually a library prepared from the sample before immunoprecipitation (input) is used as a control.

After sequencing and quality control of the resulting sequences, reads are mapped to the reference genome. The coverages of the immunoprecipitated (IP) and input samples are then compared to identify sites where the IP signal is significantly above the background noise. Some software packages take advantage of the fact that some proteins bind narrow regions of the genome to detect particularly shaped peaks of signal ([Zhang et al., 2008]).

The main limitation of ChIP-seq is its dependence on an antibody specific for the protein that is being studied. Nevertheless, it remains the de facto standard for determining protein binding sites. A recent variation, ChIP-exo, has managed to determine binding sites with single-nucleotide resolution ([Rhee and Pugh, 2011]).

Measuring occupancy genome-wide

There are other techniques to determine protein-DNA interactions in chromatin that do not have the protein specificity of ChIP-seq. One of them is MNase-seq, which is used to determine nucleosome positioning...
in a population of cells (Zhang and Pugh, 2011). Instead of using an antibody to pulldown a protein of interest, MNase-seq relies on DNA digestion by micrococcal nuclease (MNase) to digest DNA that is not wrapped around nucleosomes. The remaining DNA fragments are then sequenced and the locations of the genome to which the sequenced reads map are where nucleosomes were present in the population of cells.

Other genome-wide techniques take the inverse approach of MNase-seq: they detect regions of the genome that are depleted for nucleosomes. Formaldehyde assisted isolation of regulatory elements (FAIRE) utilises formaldehyde as a cross linking agent to bind proteins to DNA (Giresi et al., 2007). The cross-linked sample is then sonicated and sheared into small fragments by sonication. Finally, the sonicated sample is extracted by phenol-chloroform. This creates a separation between the DNA fragments that are cross-linked with nucleosome and those that are not and allows for the selection of the latter. After sequencing and mapping to the genome, the areas of the genome that were not occupied by nucleosomes are identified.

DNase-seq also identifies areas of the genome that are not occluded by proteins (Boyle et al., 2008). DNA is digested with a small amount of deoxyribonuclease I (DNase I) which will cleave DNA in those sites that are more sensitive to the enzyme due to not being protected by binding proteins. Biotinylated linkers are attached to the ends created by DNase digestion. These linkers are then used to extract DNA sequences that when mapped to the genome identify the DNase hypersensitive sites.

MNase-seq, FAIRE and DNase-seq allow for an unbiased mapping of the accessibility and occupancy of genomic DNA. This makes them important techniques in mapping the primary organisation of chromatin.

A return to colour

As sequencing becomes cheaper and more wide-spread, the number of available genome-wide datasets increases. New computational techniques bring together different proteins to segment the genome into different “colours” or states (Filion et al., 2010; Ernst et al., 2011). By using dimensionality-reduction mathematics like principal component analysis (PCA) or algorithms such as hidden Markov models (HMMs), computational biologists can characterise areas of the genome according to their protein content and accessibility. These reveal the extent to which proteins potentially co-bind in the genome and help identify different types of repressive or active chromatin based on what effectors define each state. Genome segmentation analysis goes beyond the determination of binding sites, links the individual chromatin components together and helps identify their functional relationships.
1.6.3 Capturing conformation

The publication of the chromosome conformation capture technique (3C; Dekker et al., 2002) opened the door to the determination of long-distance DNA contacts. Variations of the original technique have extended 3C, so that now the DNA-DNA contacts across the whole genome can be measured in a single experiment. The full implications of genome-wide interaction maps to genomic research are still being discovered. A detailed explanation of these techniques is presented in chapter 5.

1.6.4 Making sense of data

In recent years the size and complexity of genome-wide assays has made computational techniques essential to their analyses. A single ChIP-seq experiment might produce tens of millions of reads, plus a similar number of reads for the non-immunoprecipitated control. To convert the gigabytes of data into meaningful biological informations the scientific community has developed software tools to help deal with the different aspects of data processing, including dedicated sequence aligners to efficiently map individual reads to reference genomes. Examples of these are BWA (Zhang et al., 2008) and Bowtie (Langmead et al., 2009).

What software to use will depend on the nature of the experiment and the study being conducted. For ChIP-seq experiments one can use peak callers like MACS (Zhang et al., 2008) or PeakSeq (Rozowsky et al., 2009) to identify the protein binding sites. Much of the downstream analysis and comparison of datasets depends on ad hoc scripts created by the computational biologist analysing the data. There are projects that maintain the source code and documentation of software packages produced by the genomics community to assist in the processing, manipulation and visualisation of genomics data in a variety of programming languages. The most prominent programming language in genomics at the moment is R and the project that maintains the software for genomic data processing is Bioconductor (Gentleman et al., 2004).

In addition to the software packages used for analysis, the data itself is shared by the community as a whole. Journals enforce the deposition of datasets in repositories where any researcher — or indeed any member of the public — can download the raw and processed data. ArrayExpress (Rustici et al., 2013) and Gene Expression Omnibus (Barrett et al., 2013) are the two repositories used by the community to deposit and share their data.

Most of the datasets available are from individual papers and projects. Recently large consortia have been created with the express purpose of characterising the chromatin landscape of model organisms. These organism include Drosophila melanogaster and Caenorhabditis elegans.
in modENCODE (Celniker et al., 2009), *Mus musculus* in mouse ENCODE (Mouse ENCODE Consortium et al., 2012) and *Homo sapiens* in ENCODE (ENCODE Project Consortium et al., 2012). For each model organism several tissues, cell lines and developmental stages are extensively characterised in terms of the binding sites of their chromatin proteins, their expression patterns and the DNA accessibility. All data is made publicly available to the genomics community for use in their research projects.

1.7 STATEMENT OF OBJECTIVES

In this thesis I describe several ways in which interactions inside the nucleus lead to the regulation of transcription. These interactions are between proteins and DNA, proteins and RNA and between DNA itself. A brief description of each chapter follows.

1.7.1 *The many lives of Msl1*

Chapter 2 describes a project done in collaboration with the Akhtar laboratory from the Max Planck Institute for Immunobiology and Epigenetics. We investigated the role of Msl1 and Mof in two related *Drosophila* species. In doing so, we identified a novel genome-wide transcription-associated activity for Msl1, pointing to a possible conserved function in higher metazoans.

1.7.2 *lncRNAs in Drosophila dosage compensation*

In chapter 3, also a result of a collaboration with the Akhtar laboratory, we characterised the interaction of DCC proteins with the RNA components of the complex, roX1 and roX2, using individual-nucleotide resolution cross linking and immunoprecipitation (iCLIP) datasets. We show how the secondary structure of the RNA is intrinsically linked to the redundant behaviour of roX1 and roX2.

1.7.3 *Pol II distinct modes of action*

Chapter 4 describes a purely computational project. Using datasets generated by high-resolution chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-Seq) I identified five distinct modes of occupancy of Pol II in gene promoters in *Drosophila melanogaster*. I present evidence that these modes reflect the different transcriptional patterns of house-keeping and developmentally regulated genes. By further integrating datasets of transcription factor-binding, histone modifications, chromatin states or colours, nucleosome
organisation and gene ontology annotations, I attempt to explain the factors underlying Pol II’s different patterns of occupancy.

1.7.4 Promoter capture leads to new insights in the genome 3D architecture

Chapter 5 describes a project done in collaboration with the Fraser and Cameron laboratories of the Babraham Institute. Using datasets describing the genomic loci that interact with promoters I investigated the chromatin landscape in *Mus musculus* and *Homo sapiens*. I report a clear enrichment of transcription-related chromatin marks in regions interacting with highly-expressed promoters. These results will help elucidate the complex interplay between transcription regulation and three-dimensional genome organisation.
The work presented in this chapter was done as part of a collaboration between the Luscombe Laboratory and the Akhtar Laboratory of the Max Plank Institute for Immunobiology and Epigenetics in Freiburg. The ChIP-seq experiments and expression profiling in *Drosophila melanogaster* were performed by Thomas Conrad of the Akhtar Laboratory. All other experiments were performed by Sarantis Chlamydas, also of the Akhtar Laboratory. All data analysis was performed by me.

2.1 INTRODUCTION

2.1.1 Dosage compensation in *D. melanogaster* is achieved by male X up-regulation

As described in the introduction, organisms that show a disparity in the number of sexual chromosomes between genders have evolved mechanisms to compensate for differences in gene dosage. In *Drosophila melanogaster* the male flies have only one copy of the X chromosome, compared with two copies in females. The compensation is achieved in male cells, in which the single X chromosome undergoes double the amount of transcription compared with each female X chromosome. This way, both male and female cells produce equal amounts of mRNA for X-chromosomal genes.

The Dosage Compensation Complex (DCC) is responsible for establishing the conditions that lead to X up-regulation in males. This complex includes the proteins Male Specific Lethal 1, 2 and 3 (Msl1, Msl2, Msl3), maleless (Mle) and Males absent On the First generation (Mof), as well as one of the two non-coding RNAs roX1 and roX2.

2.1.2 Mof is an acetyl-transferase responsible for H4K16Ac

Of the proteins belonging to the DCC, Mof was the one shown to be the acetyl-transferase that acetylates H4K16 ([Akhtar and Becker, 2000](#)). Together with the other members of the DCC, Mof binds extensively to the X chromosome, but it also binds autosomal promoters ([Kind et al., 2008](#)).

The up-regulation of the male X-linked genes requires the acetylation of lysine 16 on the histone H4 subunit across much of the X chromosome (H4K16ac; [Bone et al., 1994](#)). This acetylation is thought to facilitate higher transcription levels by increasing the accessibility of chromatin
The other life of the *Drosophila* DCC

Bell *et al.* (2010). However, a recent study in mouse embryonic stem cells showed no change in chromatin compaction when this modification was lost (Taylor *et al.*, 2013), suggesting other mechanisms of H4K16ac action.

Despite Mof’s clear function in acetylating the male X chromosome, it has functions beyond dosage compensation. For instance, Mof also interacts with another protein complex, the non-specific lethal (NSL) complex (Mendjan *et al.*, 2006).

The NSL complex binds to active promoters, where it recruits Mof (Raja *et al.*, 2010). NSL target genes are transcriptionally active across a range of cell types, suggesting that the Mof-NSL complex regulates house-keeping genes in *D. melanogaster* (Lam *et al.*, 2012).

2.1.3 Msl1 directs the DCC to the X chromosome HAS

Msl1, another member of the DCC, has no known globular domains in its structure. Recent structural studies have shown that it forms dimers, and that only the Msl1 dimer interacts with Msl2, an interaction necessary for the formation of the DCC (Hallacli *et al.*, 2012). The same study found that Msl1 binds a small number of tested autosomal promoters in addition to the widespread binding on the X chromosome. When dimerisation is prevented, Msl1 no longer binds the X chromosome, but maintains promoter binding. Hallacli *et al.* (2012) suggest that Msl1 may have two modes of binding that confer different regulatory functions.

According to the current model of dosage compensation in fly, Msl1 binds a set of over 150 High-Affinity Sites on the X chromosome (Straub *et al.*, 2008), recruiting the rest of the complex with it. The DCC then binds the regions cis-adjacent to the HAS, spreading across the X chromosome from these sites. Mof’s physical interactions with Msl1 and Msl3 are essential for proper targeting to the X chromosome (Morales *et al.*, 2004; Kadlec *et al.*, 2011).

2.1.4 DCC members are present in other metazoans with very different dosage compensation mechanisms

The proteins of the DCC have mammalian orthologs, in which sexual dosage compensation is achieved using a completely different mechanism that does not involve the *D. melanogaster* DCC proteins (Marin, 2003). Them being present in mammals thus raises questions about the functions they perform. If they are active in mammals, they must have other functions besides dosage compensation.

While these proteins were initially studied for their relationship with dosage compensation in *D. melanogaster*, we now know that Mof has a genome-wide function as a regulator of the transcription of house-keeping genes. An analysis of its functional domains indicates that
the domains necessary for dosage compensation are *Drosophila*-specific, suggesting that Mof was co-opted for dosage compensation in flies (Conrad *et al.* 2012a).

As for Ms1, the promoter binding observed in Hallaci *et al.* (2012) suggests it has functions beyond dosage compensation, but these have not yet been characterised. Knowing what these are will help shape our understanding of possible conserved behaviours of Ms1.

### 2.1.5 Objective

In order to identify conserved functions of DCC proteins, we studied the binding behaviours of Mof and Ms1 in two fly species: the model organism *D. melanogaster* and another member of the genus, *Drosophila virilis*. These two *Drosophila* species diverged around 40 million years ago. By assaying the behaviour of DCC proteins in different species we are able to identify their conserved functions. We find that Ms1 binds promoters genome-wide as well as on the X chromosome. We then proceed to investigate whether this behaviour is linked to the presence of Mof and the mechanism through which this additional binding operates.

### 2.2 Results

#### 2.2.1 Mof binding in *D. melanogaster* males

To characterise Mof binding in *D. melanogaster* I reprocessed the data from Conrad *et al.* (2012a) to reproduce the results and obtain a set of Mof-bound genes with which to compare the Mof behaviour in *D. virilis*. The samples were obtained from third-instar male larvae. I mapped the raw reads of the immunoprecipitated sample (IP) and input control the *D. melanogaster* genome using Bowtie (Langmead *et al.* 2009). Bowtie 1 was set to the default parameters with the exception that only the best alignment for each read was reported. The IP sample had 12,205,167 reads, of which 4,628,406 (37.9%) mapped to the *D. melanogaster* genome. This discrepancy might be due to a less efficient immunoprecipitation. The input control a total of 10,151,169 reads were sequenced, 9,562,355 of which (94.2%) mapped to the reference genome.

To identify the bound regions from the mapped reads, I used the approach described in Conrad *et al.* (2012a). The genome was divided into 25 bp bins and the IP and input reads that fall within each bin were counted. Using DESeq (Anders *et al.* 2013), I normalised the bin counts for the different number of mapped reads in each sample and calculated the log2 fold-change of IP over input. The log2 fold-change values were smoothed using the mean of a sliding window with a width of 400 bp. To determine the threshold for binding, I took the log2 fold-
change values under the mode of the distribution and reversed it to create a symmetric null distribution. A threshold for a FDR-adjusted p-value cut-off of 0.05 was then identified. Bins below this cut-off were considered significantly bound in the following analysis.

In total, over 5.6 megabase pairs (Mb) of the genome are bound by Mof. The bound regions form 29,565 contiguous stretches with a mean width of 114.7 bp in autosomes and 246.8 bp in the X chromosome. To observe how Mof binds genes, I took the window within 2000 bp of each transcription start site (TSS) and the window 2000 bp upstream of the transcription termination site (TTS) for each gene and averaged the log2 fold-change profiles for those positions. The resulting values are shown in figure 2.1 for all genes (black line), for autosomal genes (blue line) and for X-linked genes (red line). It is clear that there is a peak of binding just before the TSS in genes across the entire genome and that X-linked genes show strong binding along the gene body in addition to the TSS.

To identify a set of promoter-bound genes, I selected those with a significantly bound bin in the area around the average peak, which is between 200bp upstream of the TSS and the TSS. This area is marked by grey dashed lines in figure 2.1. In the gene body, the distribution of binding is bimodal, with genes showing either almost no binding or complete binding across exon bins starting 500 bp downstream of the TSS. Therefore, to select body-bound genes, I identified those with
more than 60% of their bins overlapping exons classified as significantly bound.

Table 2 shows the number of X-linked and autosomal genes bound by Mof in the promoter and in the gene body. As suggested by the average gene profile, both X-linked and autosomal genes are bound at promoters (60.7% and 28.9%, respectively). In contrast, almost all body-bound genes are present in the X chromosome, where 43.2% of this chromosome’s genes are in this category.

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Autosomes</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Promoter-bound</td>
<td>4934</td>
<td>3552</td>
<td>1382</td>
</tr>
<tr>
<td></td>
<td>(33.8%)</td>
<td>(28.9%)</td>
<td>(60.7%)</td>
</tr>
<tr>
<td>Gene body-bound</td>
<td>998</td>
<td>14</td>
<td>984</td>
</tr>
<tr>
<td></td>
<td>(6.8%)</td>
<td>(0.1%)</td>
<td>(43.2%)</td>
</tr>
</tbody>
</table>

Table 2: Number of genes bound by Mof in D. melanogaster

The average gene profiles in figure 2.1 include all genes, including those that are classified as not being bound by Mof. To see if the same profiles are visible when considering only those genes with which Mof interacts, I recreated the plot in figure 2.1b using only promoter-bound genes (full lines) and unbound genes (dashed-lines). This allows for a comparison of the binding level across the population in bound genes. Apart from an overall increase in the log2 fold-change values of the average profiles in the bound genes, the profiles are similar to the previous plot. Mof in autosomes has a sharp peak at the promoter and in the X chromosome it binds both at the promoter and inside the gene body.

2.2.2 Mof binding in D. virilis males

For the analysis of Mof binding in D. virilis, we created a ChIP-seq dataset using whole third-instar male larvae. The processing was done in the same way as for the D. melanogaster Mof samples, but using the D. virilis genome as a reference. The IP sample had 32,292,847 reads sequences, while the input had 36,460,591, of which 30,047,606 (93.0%) and 35,319,561 (96.9%), respectively, map to the D. virilis genome.

More than 20.3 Mb of the D. virilis genome are bound by Mof, in more than 46 thousand contiguous regions. As in D. melanogaster, the mean length of the bound regions is longer in the X chromosome (477.8 bp) than in autosomes (385.3 bp). The increase in length of the bound regions compared to D. melanogaster may be due to the increased number of reads available for the D. virilis sample. Mof also behaves similarly to D. melanogaster at TSSs and gene bodies. As seen in figure 2.2, there is a peak upstream of the TSS in both X-linked and autosomal genes, whereas binding over the gene body is only present in X-linked genes. As the peak before the TSS is slightly shifted upstream compared to D. melanogaster, in D. virilis I identified promoter-bound genes by selecting those with a significantly bound bin in a window
between 375 bp and 175 bp upstream of the TSS (marked in Figure 2.2 by black dashed vertical lines). The shift in the peak of Mof binding might be an actual biological difference or it might be due to a worse annotation of the start coordinates of D. virilis genes. Cap analysis gene expression (CAGE) datasets for D. virilis would allow for a more precise mapping of the gene start coordinates.

![Figure 2.2](image)

Figure 2.2: (a) Average gene profiles of Mof binding in male D. virilis larvae for all (black line), autosomal (blue line) and X-linked genes (red line). The average log2 fold-change values of IP over input were calculated for the positions within 2000 bp of the TSS and for the positions 2000 bp upstream of the TTS. (b) Average gene profile for the subset of genes bound by Mof at the promoter (solid line) and for those not bound (dashed line).

The more extensive binding in D. virilis as a whole is reflected in the number of genes bound, but the overall differences between the X chromosome and autosomes are still observed. As table 3 describes, a very high proportion of X-linked genes (72.6%) is promoter-bound. Autosomal genes show a smaller proportion of bound genes, but it is still close to 40%. As in D. melanogaster, for gene body-bound genes the discrepancy between autosomes and X-linked genes is much more pronounced, with just 5.7% of the former being bound compared to 54.0% of the latter.

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Autosomes</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Promoter-bound</td>
<td>6727</td>
<td>4842 (39.7%)</td>
<td>1885 (73.6%)</td>
</tr>
<tr>
<td>Gene body-bound</td>
<td>2082</td>
<td>699 (5.7%)</td>
<td>1383 (54.0%)</td>
</tr>
</tbody>
</table>

Table 3: Number of genes bound by Mof in D. virilis

The average profiles of the promoter-bound genes (Figure 2.2b) show that the peak at autosomal promoters is more pronounced than the average from all autosomes would suggest. Despite an increase in the
number of bound genes, the behaviour of Mof in *D. virilis* is very similar to that in *D. melanogaster*: Mof binds promoters throughout the genome and binds gene bodies on the X chromosome.

### 2.2.3 Msl1 binding in *D. melanogaster* males

We also characterised the behaviour of Msl1, another component of the DCC. We created ChIP-seq libraries from third-instar male *D. melanogaster* larvae, which were processed in the same way as the Mof samples. As this sample was from the same chromatin preparation as the sample used for Mof in *D. melanogaster*, the same input was used. The Msl1 IP sample gave 19,577,660 reads in total, of which 10,733,235 (54.8%) mapped to the reference genome.

![Figure 2.3](image)

**Figure 2.3:** (a) Average gene profiles of Msl1 binding in male *D. melanogaster* larvae for all (black line), autosomal (blue line) and X-linked genes (red line). The average log2 fold-change values of IP over input were calculated for the positions within 2000 bp of the TSS and for the positions 2000 bp upstream of the TTS. (b) Average gene profile for the subset of genes bound by Msl1 at the promoter (solid line) and for those not bound (dashed line).

The amount of the *D. melanogaster* genome bound by Msl1 is over 1 Mb, in 6044 distinct regions. The average width of these regions is not substantially different between the autosomes (173.5 bp) and the X chromosome (179.7 bp). The average profiles show a very pronounced peak of Msl1 at the promoters of both X-linked and autosomal genes. Binding along the gene body in X-linked genes is higher than along the gene body of autosomal genes, but this difference is much smaller than what was observed for Mof.

The numbers of genes classified as bound which are presented in Table confirm the result from the average profiles. In *D. melanogaster* male larvae and using the same criteria as for Mof, Msl1 does not
bind gene bodies on either the X chromosome or autosomes (0.4% of X-linked genes compared to 0.0% of autosomal genes). Msl1 binds X-linked promoters (28.5% of all X-linked promoters) and autosomal promoters (18.1% of all autosomal promoters). When selecting only the promoter-bound genes for plotting, the peak of Msl1 at the promoter in autosomal genes becomes even more pronounced (Figure 2.3b). That Msl1 binds autosomal promoters as well as those on the X chromosome has since been reported in another study (Straub et al., 2012), and may reveal a new function for this hitherto DCC-exclusive protein.

2.2.4 Msl1 binding in D. virilis males

We also investigated Msl1 binding behaviour in D. virilis to see whether it is conserved. This ChIP-seq experiment was carried out in whole third-instar larvae and the reads obtained from it were processed in the same way as in the previous experiments to obtain the genomic locations where Msl1 binds to the D. virilis genome. The IP and input samples had a total of 306,453,011 and 328,900,256 sequenced reads, of which 177,523,455 (57.9%) and 190,340,046 (57.9%), respectively, mapped to the D. virilis genome. In this experiment the log2 fold-change values denoting the enrichment of IP over input were lower than observed in the D. virilis Mof and D. melanogaster Mof and Msl1 experiments. This lower enrichment results in no region passing the p-value threshold for binding. The lower values can be observed in Figure 2.4. However, we can also see from the average profiles that, even though the effect is diminished, there is an increased level of Msl1 at the promoters of both autosomal and X-linked genes.

2.2.5 DCC binding is conserved between orthologs

Having determined the sets of genes bound by Mof and Msl1 in the two Drosophila species, I proceeded to see the extent to which binding of the two proteins is conserved. I downloaded a set of one-to-one orthologs from the Ensembl Compara database (Kersey et al., 2012) and integrated them into the analysis.

In Figure 2.3 I present all pairs of orthologs in a heat map. Each line corresponds to a pair of orthologs and the colours represent the log2 fold-change values of Mof and Msl1 in D. melanogaster and D. virilis. The orthologs are ordered first based on whether they are on
Figure 2.4: Average gene profiles of Msl1 binding in male *D. virilis* larvae for all (black line), all autosomal (blue line) and all X-linked genes (red line). The average log2 fold-change values of IP over input were calculated for the positions within 2000 bp of the TSS and for the positions 2000 bp upstream of the TTS.

the X chromosome or on autosomes in *D. melanogaster* and, inside those groups, by decreasing intensity of Mof binding at the promoter. By looking at the intensity of log2 fold-change values for each pair of genes, we can conclude that Mof and Msl1 not only bind the same genes in *D. melanogaster*, but also that the corresponding orthologous genes in *D. virilis* tend to also have similar binding of Mof and Msl1.

*Mof gene body-binding is X-dependent*

To investigate the connection between conservation of binding and genomic location I looked at cases where the orthologs are on the X chromosome in the two species and compared this to situations where one member of the pair is on the X in one species and another is on the autosomes in the other species. This analysis was done for Mof only, to compare the conservation of the two modes of binding observed for this protein: binding at the promoter and binding over the gene body. Every pair of orthologs was classified as being bound in both species, in just *D. virilis*, in just *D. melanogaster* or in neither. This was done for both patterns of Mof binding (promoter and gene body).

In Figure 2.6a I plot the proportion of ortholog pairs where both members are on the X chromosome that are bound in both species, or in just one of them. For both promoter binding and gene body binding a high proportion of the pairs are bound in both species. The other set of orthologs analysed were those where the gene is in the X chromosome in one of the species and in an autosome in the other species. In this set only the promoter binding is conserved between the two species and gene body binding, by contrast, is present only in one of the species (Figure 2.6b).
Figure 2.5: Heat map representing the intensity values of Mof and Msl1 ChIP-seq experiments in *D. melanogaster* and *D. virilis* orthologs. Each line represents a pair of one to one orthologous genes between the two species. The colour reflects the log2 fold-change values at positions relative to the TSS and TTS of the genes.
Figure 2.6: Conservation of Mof binding between *D. melanogaster* and *D. virilis*. The pairs of orthologs where Mof was binding at the promoter (blue bars) or in the gene body (red bars) in either of the two species were selected. In (a) the proportion of pairs in which both species are bound, only *D. virilis* is bound or only *D. melanogaster* is bound is shown for the pairs where both orthologous genes are in the X chromosome. In (b) are shown the proportions for those pairs where one member is in the X chromosome and the other is in an autosome.

These results show that the two behaviours of Mof rely on different contexts. Mof binding at the promoter depends on the gene itself, as reflected by its high degree of conservation between the two species. Mof binding on the gene body, on the other hand, depends on the gene being present on the X chromosome. Even for orthologs, the gene body binding is not maintained when one of the members of the pair is located on an autosome.

### 2.2.6 Downsampling of *D. virilis* Mof dataset does not affect the balance between binding of X and autosomal genes

The ChIP-seq datasets used to compare the binding of Mof between the two *Drosophila* species in the previous sections have a sizeable disparity in the number of reads of each sample. The *D. melanogaster* sample has under ten million mapped reads, while the *D. virilis* one has over thirty million. The increased number of reads in the *D. virilis* sample can result in increased sensitivity in detecting genomic regions bound by the protein, which in turn can lead to the higher number of bound genes observed in this species compared to *D. melanogaster*.

In order to assess the extent to which the different number of reads in the two ChIP-seq experiments impacted the comparison, I performed a random downsampling of the *D. virilis* Mof IP and input samples to ten,
million mapped reads, an amount of the same order of magnitude as the 
*D. melanogaster* Mof ChIP-seq experiment. After the downsampling,
the mapped reads were processed in a manner identical to original
samples.

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Autosomes</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Promoter-bound</td>
<td>5745 (38.9%)</td>
<td>4180 (34.3%)</td>
<td>1565 (61.1%)</td>
</tr>
<tr>
<td>Gene body-bound</td>
<td>1384 (9.4%)</td>
<td>183 (1.5%)</td>
<td>1201 (46.9%)</td>
</tr>
</tbody>
</table>

Table 5: Number of genes bound by Mof in the down sampled *D. virilis* experiment

The number of genes classified as bound in the down sampled dataset are present in table 5. Compared to the non-down sampled dataset, there is an overall decrease in the number of genes bound, bringing the numbers closer to the number of bound genes by Mof in *D. melanogaster*. This decrease does not change the balance of bound genes between the X chromosome and the autosomes. Thus, the finding that Mof behaves in the same manner in the two *Drosophila* species here studied, by binding the promoters of genes in all chromosomes and the gene bodies of genes in the X chromosome, is not affected by the downsampling control. Similarly, the analysis of Mof binding in pairs of orthologs presented in figure 2.6 does not change substantially when the down sampled *D. virilis* dataset is used.

2.2.7 HAS sequence specificity is conserved

Some regions of the *D. melanogaster* X chromosome have been characterised as having a high affinity for the DCC members. These High Affinity Sites (HAS) show particularly high occupancies for the various DCC proteins in wild-type conditions and some members of the complex are bound even under conditions where the complex has been disrupted by deleting some of its components. Interestingly, HAS contain a particular GA-rich motif ([Alekseyenko et al., 2008](#)).

Previous studies of HAS relied on conditions in which DCC had been disrupted. While this is not the case for the datasets at my disposal, I can use the definition of HAS as genomic regions where the DCC proteins show a high level of binding. I took the bins that are in the top 0.1% of log2 fold-change values for both Mof and Msl1 in *D. melanogaster*. The motif from these putative HAS was found using MEME ([Bailey et al., 2006](#)) and as shown in Figure 2.7b, is very similar to the one previously described (Figure 2.7a).

Having ascertained that using the intersection of the top binding sites for Mof and Msl1 in *D. melanogaster* recovers the motif associated with HAS, I applied the same procedure to the *D. virilis* samples. The recovered motif is GA-rich and similar to the one found in *D. melanogaster*
2.2 results

(a) D. melanogaster (Alekseyenko et al., 2008)

(b) D. melanogaster

(c) D. virilis

Figure 2.7: The motif found in a previous study of HAS (a) and in the regions that show the highest intensity of binding for both Mof and Msl1 in D. melanogaster (b) and D. virilis (c).

(Figure 2.7c). This analysis shows that the sequence specificity of the DCC at its HAS is conserved between the two species.

2.2.8 Msl1 binding in the absence of dosage compensation and Maf

I showed the presence of Msl1 in promoters of not only X-linked genes, but also of autosomal promoters. The presence of Msl1 in promoters of autosomal genes suggests that this protein has a function beyond dosage compensation of the X chromosome. If this is the case, then Msl1 should also bind promoters in conditions where dosage compensation is not necessary and where the DCC does not form. We created ChIP-seq datasets for Msl1 in two such conditions: female larvae and male larvae where the C-terminal region of Mof has been deleted. This region contains the HAT domain of Mof that physically interacts with Msl1. These mutants are named mof2 (Conrad et al., 2012a). The data for both conditions was processed in the same manner as the other ChIP-seq experiments presented in this chapter.

The Msl1 ChIP-seq experiment in D. melanogaster female larvae yielded 17,762,885 for the IP sample and 15,784,565 for the input. Of these, 14,288,270 (80.4%) and 14,958,155 (94.8%), respectively, mapped to the reference genome. In females, Msl1 binds an area of more than 300 thousand base-pairs, smaller than the area bound by Msl1 in males. In the same manner, the number of individual contiguous regions bound, at 1287, is smaller than in males. These regions are on average slightly larger in the X (273 bp) than in autosomes (226.3 bp). The average profiles over all X-linked or autosomal genes show that there is a peak of Msl1 at the promoter that is higher in the former than in the latter. However, this difference is simply a reflection of the number of genes bound by Msl1, as it diminishes when only promoter-bound genes are chosen. More importantly, the genes bound by Msl1 at the promoter show lower binding values in the gene body.
of X-linked genes compared to wild-type males, indicating the lack of dosage compensation-related binding in females.

\[ \text{Table 6: Number of genes bound by Msl1 in } D. \text{ melanogaster} \text{ female larvae} \]

<table>
<thead>
<tr>
<th>Type of Binding</th>
<th>All</th>
<th>Autosomes</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Promoter-bound</td>
<td>867 (5.9%)</td>
<td>591 (4.8%)</td>
<td>276 (12.1%)</td>
</tr>
<tr>
<td>Gene body-bound</td>
<td>9 (0.1%)</td>
<td>9 (0.1%)</td>
<td>0 (0.0%)</td>
</tr>
</tbody>
</table>

The ChIP-seq experiment of Msl1 in \textit{mof2} mutant male larvae resulted in 11,003,368 reads sequenced from the IP sample, 6,166,451 (56.0%) of which map to the reference genome and 11,307,076 reads sequenced from the input, 10,455,587 (92.5%) of which map to the genome. In the \textit{mof2} mutant male flies, the area bound by Msl1 is above three Mb, divided over 20032 contiguous regions. Under these conditions, the regions bound have a mean width of 137.3 bp in the X chromosome and of 172.9 bp in the autosomes. The average gene profile, represented in Figure 2.9, maintains the peak at the promoter and shows higher values of Msl1 in the body of X-linked genes compared to autosomes, similar to the profile seen in the wild type males.
The observation that Msl1 binds promoters on the autosomes of male wild type flies suggests that the protein plays a role beyond dosage compensation. The results from the moft2 mutant male larvae and from the wild type female larvae confirm that this behaviour also occurs in situations where the dosage compensation is either defective (the former) or absent altogether (the latter).

Table 7: Number of genes bound by Msl1 in D. melanogaster moft2 mutant larvae

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Autosomes</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Promoter-bound</td>
<td>5391 (40.0%)</td>
<td>4214 (34.2%)</td>
<td>1177 (51.7%)</td>
</tr>
<tr>
<td>Gene body-bound</td>
<td>62 (0.4%)</td>
<td>16 (0.1%)</td>
<td>46 (2.0%)</td>
</tr>
</tbody>
</table>

When the genes that are bound are quantified in Table 7, it becomes clear that the increase in signal in the body of X-linked genes is still below the significance threshold. So, using the strict criteria, the number of body-bound genes is negligible both in the X chromosome and in autosomes. We observe promoter-binding both in the X chromosome and in the autosomes, where 51.7% and 34.2% of the genes there are bound, respectively.

Figure 2.9: (a) Average gene profiles of Msl1 binding in male D. melanogaster larvae that carry a truncated version of Moft (moft2) for all (black line), all autosomal (blue line) and all X-linked genes (red line). The average log2 fold-change values of IP over input were calculated for the positions within 2000 bp of the TSS and for the positions 2000 bp upstream of the TTS. (b) Average gene profile the genes bound by Msl1 at the promoter (solid line) and for those not bound (dashed line).
2.2.9 Msl1 binding is associated with higher expression

Next we examined whether Msl1 binding has an effect in the transcriptional activity of bound genes. We used expression profiling data using GeneChip arrays from wild type and mof2 mutant male larvae. The expression profiling was done in triplicates for wild type and mutant larvae. I normalised the intensity values across the different arrays using the GCRMA Bioconductor package, and assigned the average probeset intensity as the expression level of the genes to which they map. Afterwards, I assigned the genes to two groups, X-linked and autosomal and, inside each of these groups, subdivided according to whether they are not bound by Msl1, bound at the promoter, or bound at the gene body.

In Figure 2.10a, I plot the expression levels of the genes from the different location and Msl1 binding categories for wild type males. The number of genes considered in each category are displayed above the box plot labels. For the X-linked genes, we can see that the highest expression is that of the genes that have Msl1 in the gene body. However, there is also a very clear difference between the promoter-bound genes and those that are not bound, with the former showing higher expression. This difference between promoter-bound and not bound genes is also observed in autosomes.

When a similar analysis is done in Figure 2.10b for the mof2 mutant male flies, the increased expression of Msl1-bound genes is also present. In this condition, Mof has been truncated and has lost its acetyl-transferase function, suggesting that the increased expression of Msl1-bound genes is independent of histone acetylation. However, there might still be some Mof protein present in the cells or some residual maternal acetylation.

To confirm that histone acetylation plays no role in the increased expression of Msl1-bound genes, I used a list of genomic sites that are acetylated in the lysine 16 of the histone 4 (the mark deposited by Mof) from third-instar larvae salivary glands in both the wild type and mof2 mutant conditions to classify the promoters as being acetylated or not. This list was provided by Florence Cavalli of the Luscombe Lab. Figures 2.10c and d show the expression values for acetylated and non-acetylated promoters in the wild type and mutant flies. In both cases, we see an increased expression of non-acetylated Msl1-bound genes compared to those that are also non-acetylated but that are not bound by Msl1.

By integrating the Msl1 ChIP-seq results with expression profiling, we identified an association between the genome-wide promoter binding of Msl1 and higher expression values. By further adding information about which promoters are acetylated, we conclude that this association is independent of Mof activity.
2.2 Results

Figure 2.10: Expression levels in wild type and mof2 male larvae of genes that are bound or not by Msl1. Under both conditions and both for X-linked and autosomal promoters Msl1-bound genes show higher expression levels than non-bound genes. Expression levels in wild type and mof2 male larvae of genes that are bound or not by Msl1 and that have, or not H4K16ac in their promoters. The higher expression of Msl1-bound genes is not dependent on histone acetylation.
2.2.10  *Msl1 depletion leads to reduction of Pol II Ser5 levels*

The results show that Msl1 binds promoters on both X chromosome and autosomes and that this binding is associated with higher expression. To understand the mechanism through which Msl1 acts, we carried out Msl1 RNAi experiments. We can thus see the impact of Msl1 deficiency in cells. The RNAi experiments were performed in both S2 cells, which have a male phenotype, and Kc cells, with a female phenotype. As a control, GFP RNAi experiments were also carried out.

We used Western blots to measure the presence of members of the DCC, non-specific lethal 3 (NSL3), TATA-box binding protein (TBP), the RNA Polymerase II (Pol II) subunit RPB3, Pol II phosphorylated in the serine 5 (PolII-Ser5) and serine 2 (PolII-Ser2) of its C-terminal domain and CDK7, a kinase responsible for the phosphorylation of Ser5. We also assayed several histones and histone modifications: H4K16ac, histone H3 in bulk and in its lysine 4 trimethylated form (H3K4me3) and histone H2B in bulk and in its lysine 120 ubiquitinated form (H2BK120ub).

The results of the Western blots are presented in Figure 2.11a for S2 cells and 2.11b for Kc cells, with the PolII-related data highlighted with red boxes. In both figures, lane 1 and 2 correspond to the GFP RNAi, with 1 having the lower values of GFP, and lanes 3 and 4 correspond to the Msl1 RNAi, with 3 having the lower values of Msl1. Apart from Msl1 itself, the main difference between the Msl1 RNAi and the control GFP RNAi is a decrease of PolII-Ser5 when Msl1 is depleted in cells. The levels of H4K16ac and H2BK120ub are also decreased upon Msl1 depletion. The reduction in H4K16ac levels might be a side-effect of the lack of Msl1 disrupting the targeting of the DCC complex. As for H2BK120ub, the reduced levels of PolII-Ser5 might diminish the cascade that leads to the ubiquitination of H2B.

Following up on the depletion of PolII-Ser5 after Msl1 RNAi, we performed ChIP experiments for PolII-Ser5 on four Msl1-bound genes. The promoter, middle and end of the genes were probed, both in Msl1 RNAi and GFP RNAi conditions. Figure 2.12a shows the levels of recovered input in the ChIP experiments. In the control GFP RNAi conditions, PolII-Ser5 has higher levels at the promoter of the four genes. These levels are greatly reduced in the Msl1 RNAi.

We also performed ChIP experiments in the same conditions and in the same locations for CDK7, the kinase responsible for PolII’s serine 5 phosphorylation. Consistent with the results for PolII-Ser5, upon depletion of Msl1, we observe lower levels of CDK7 in tested genes.

By depleting the levels of Msl1 in the cells, we are able to probe the mechanism by which it functions. We link Msl1 presence to the phosphorylation of Ser5 of the PolII CTD and show that its absence leads to diminished recruitment of CDK7 to Msl1-targeted genes.
Western blots to detect the levels of proteins in cells under both Msl1 RNAi and in a control GFP RNAi. Lanes 1 and 2 correspond to the GFP RNAi experiment. Lanes 3 and 4 correspond to the Msl1 RNAi experiment. Lanes 2 and 4 have a higher loading of protein concentration compared to lanes 1 and 3, respectively. The experiments were carried out in S2 (a) and Kc (b) cells.

Figure 2.11: Western blots to detect the levels of proteins in cells under both Msl1 RNAi and in a control GFP RNAi. Lanes 1 and 2 correspond to the GFP RNAi experiment. Lanes 3 and 4 correspond to the Msl1 RNAi experiment. Lanes 2 and 4 have a higher loading of protein concentration compared to lanes 1 and 3, respectively. The experiments were carried out in S2 (a) and Kc (b) cells.
2.2.11 CDK7 inhibition alters Msl1 localization

As Msl1 alters CDK7 activity, we tested whether the inverse is true. To this end, we used immunostaining of Msl1 and PolII-Ser5 in polytene chromosomes in the presence and absence of BS-181, an inhibitor of CDK7 (Ali et al., 2009). The results from this experiment are in Figure 2.13. In the top row we see the results in the absence of BS-181 and in the bottom in its presence. When CDK7 is inhibited, both PolII-Ser5 and Msl1 show heavily disrupted binding localisation compared to the control.
2.3 Discussion

Mammals compensate for the disparity in the number of sexual chromosomes between males and females using a completely different system from flies, yet proteins that are known to perform dosage compensation in flies are still present in mice and humans. In this work we assayed the genomic localisation of Mof and Msl1, two members of the DCC, in male larvae of *D. melanogaster* and *D. virilis*, two fly species that diverged roughly 40 million years ago. In doing so, we hoped to identify conserved behaviours of these proteins that would shed light on the function that they have in mammals.

In comparing the behaviour of Mof in both fly species, we see that its binding to gene bodies relies on the chromosomal context. Mof is also present at promoters, as part of the Mof-NSL context that regulates housekeeping genes ([Lam et al., 2012](#)). Not being connected to sexual chromosomes, the conservation of this behaviour between the two species is dependent only on the orthology of the genes involved. Recent studies in mouse embryonic stem cells have shown Mof to localise to promoters as well ([Taylor et al., 2013](#)). As an extension to the work presented here it would be useful to see if the conservation of Mof binding present in the two flies species studied extends to mammals. If that is the case, the dosage compensation-independent activity of this acetyl-transferase would reflect a highly conserved regulator of transcription. Interestingly, the same study also found that H4K16ac and Mof mark a novel type of enhancer, independent of the well known enhancer regulator p300. While the denser *D. melanogaster* genome makes it harder to identify distant regulatory regions, it will be interesting if H4K16ac has such a function in flies, in which case the power of the fly as a model organism could be used to further elucidate the workings of this new regulatory element.

Msl1 was only known to act as part of the dosage compensation mechanism, by binding extensively to the X chromosome. Localised ChIP experiments, however, showed it binding to some promoters of autosomes and X-linked genes, even when the dimerisation necessary for proper functioning of the DCC did not take place ([Hallachi et al., 2012](#)). We found that Msl1 binds promoters of both X-linked and autosomal promoters genome-wide in both fly species. This finding was novel at the time it was made, but has since been reported in other studies ([Straub et al., 2012](#)). Intrigued by this genome-wide behaviour of Msl1 in male flies, we inquired whether it is maintained in *D. melanogaster* under conditions where dosage compensation is not necessary or has been disabled. This is indeed the case, as Msl1 binds promoters genome-wide in both wild type female larvae and in male larvae that carry a truncated version of Mof. Using expression profile of wild type and mutant males, we also linked Msl1 promoter-binding to higher expression levels.
A serious caveat of the work presented in this chapter is the absence of replicates for the ChIP-seq experiments. The immunoprecipitation step of ChIP-seq is very sensitive to the quality of the antibody used and can have low efficiency. Depending on the immunoprecipitation efficiency, a higher proportion of sequenced reads may come from non-bound regions, resulting in a lower signal to noise ratio of the experiment and a decrease in the sensitivity and specificity of the peaks detected. Therefore replicated ChIP-seq experiments allow for a better identification of the real DNA binding patterns of the proteins assayed. The results based on the non-replicated ChIP-seq experiments presented here should be confirmed by performing the necessary replicates and repeating the analysis. The ChIP-seq of Msl1 in *D. virilis* in particular should be repeated, as the low signal despite the vast number of mapped indicates a very inefficient immunoprecipitation.

Another issue in the analysis of the ChIP-seq results is the disparity in the number of reads from the different samples, especially the difference between the *D. melanogaster* and *D. virilis* samples. An increase in the sequencing depth of a ChIP-seq library increases the sensitivity of the experiment and allows for the detection of a higher number of bound regions. The increased number of genes bound by Mof in *D. virilis* compared to *D. melanogaster* might be influenced by the disparity in the number of mapped reads. It becomes important to downsample the sample with the higher number of reads to better compare the two species. While the down sampled *D. virilis* Mof sample still shows a higher proportion of bound genes compared to the *D. melanogaster* Mof sample, the effect is diminished from the full sample.

Knowing that Msl1 binds to promoters does not by itself elucidate the mechanism by which it acts. In a set of small-scale experiments, we linked Msl1 depletion to lower cellular levels of PolII-Ser5 and to reduced levels of PolII-Ser5 and CDK7 to Msl1-targeted genes. CDK7 is the kinase that is responsible for this specific PolII CTD phosphorylation. The CTD of PolII is phosphorylated at Ser5 during initiation of transcription and is responsible for the proper capping of nascent mRNA. We hypothesise that Msl1 helps recruit CDK7 to the promoter of active genes. Msl1 promoter-binding is conserved between the fly species studied; it could also be conserved in mammals, which have a Msl1 ortholog.

In this work we extend our understanding of Mof and Msl1. These two proteins were first characterised and studied as part of their behaviour in dosage compensation in flies. While this is a very important model system for large-scale regulation of transcription, the specific mechanism is restricted to flies. We know now that Mof functions in other processes, and show here that the same is true for Msl1. The knowledge here gained will shed a light into the functions that these conserved proteins have in mammals such as ourselves.
ICLIP REVEALS THE PROTEIN-RNA INTERACTION OF THE DOSAGE COMPENSATION COMPLEX

The work presented in this chapter is a collaboration between researchers from several laboratories. Ibrahim Ilik from the Akhtar Laboratory performed the Chip-Seq, iCLIP and ChIRP experiments, Plamen Georgiev from the same laboratory carried out the viability assay, Jeffrey Quinn from the Chang Laboratory did the SHAPE experiments and analysis and I processed and analysed the ChIP-seq and iCLIP experiments.

The results from this collaboration were published in [Ilik et al. 2013].

3.1 INTRODUCTION

There are many forms of non-coding RNAs, such as micro RNAs and short interfering RNAs. A specific class of RNAs are long non-coding RNAs (lncRNAs), which are non-protein coding transcripts longer than 200 bp [Maenner et al. 2012]. In mammals there are thousands of lncRNAs, and they have been implicated in different regulatory processes. For example, HOTAIR is linked to the developmentally important HOX genes [Rinn et al. 2007] and was shown to act as a scaffold that bridges together two different chromatin complexes [Tsai et al. 2010]. Other lncRNAs have been suggested to have enhancer-like functions in human cells [Orom et al. 2010]. A third example of lncRNA action is in mammalian dosage compensation, where one of the two female X chromosomes is inactivated and a major role is played by a 17 kb long RNA, Xist.

lncRNAs are also present in *Drosophila melanogaster*. The bithorax complex, for example, contains non-coding RNAs [Lipshitz et al. 1987; Sanchez-Herrero and Akam, 1989; Cumberledge et al. 1990]. These non-coding RNAs originate at Polycomb group response elements and their transcription was shown to prevent silencing by the Polycomb group complexes in a reporter gene system [Schmitt, 2005]. However, work in fly embryos indicates that the transcription of the bithorax complex non-coding RNAs is associated with repression of transcription [Petruk et al. 2006]. A more recently characterised gene, CASK regulatory gene (CRG), expresses a long non-coding RNA in the fly central nervous system and regulates a neighbouring gene [Li et al. 2012]. Despite the fact that, in contrast to mammals, dosage compensation in *D. melanogaster* relies on the upregulation of the male X chromosome, *D. melanogaster* dosage compensation also uses lncRNAs.
In chapter 2, I explored the interactions between two proteins of the dosage compensation complex (DCC), Male Specific Lethal 1 (Msl1) and Males absent On the First (Mof), and DNA. However, there is more to the DCC than DNA and proteins. It has long been known that two lncRNAs are part of the complex. These are called, appropriately, RNA on the X 1 and 2 (roX1 and roX2, respectively). Also in collaboration with the Akhtar Laboratory, I investigated the interaction between the roX RNAs and two protein members of the complex. These proteins were Maleless (Mle), a DNA/RNA helicase, and Male Specific Lethal 2 (Msl2), a protein with a zinc finger domain.

3.1.1 The D. melanogaster DCC and RNA

roX1 and roX2 are integral parts of the DCC

In 1997, Humbert Amrein and Richard Axel created cDNA libraries from mushroom body cells from male and female flies to identify genes that differ in expression pattern between the two sexes. They succeeded in identifying two RNAs without a coding frame, which they named roX1 and roX2. The genes coding for both RNAs reside on the X chromosome. The main RNA species of roX1 and roX2 differ greatly in length, with the roX1 RNA being 3.7 kb long, while the roX2 transcript, at 1.2 kb, is less than half as long.

The two genes were then shown to be coexpressed and their transcripts to colocalize to the male X chromosome with the Male-Specific Lethal proteins (Franke and Baker, 1999; Meller et al., 2000). Other studies showed that in S2 cells, which only express roX2, this RNA is precipitates in anti-Msl1 and anti-Mle pulldowns, along with the other protein members of the complex (Smith et al., 2000; Meller et al., 2000). These results established the roX RNAs as functional, integral parts of the DCC. A more recent study used Chromatin Isolation by RNA Purification (ChIRP) to determine more precisely where in the X chromosome roX2 binding occurs and found that it has a preference for the 3’ end of genes (Chu et al., 2011).

The genomic loci encoding for both roX RNAs are among the first High-Affinity Sites (HAS) for the dosage compensation complex to be described (Kelley et al., 1999). These are sites on the X chromosome where partial complexes are observed even in conditions that prevent the proper assembly of the DCC. HAS are postulated to act as the sites from which the DCC spreads along the chromosome. Thus roX RNAs have a double function in dosage compensation: their transcripts as members of the DCC and their genomic loci as HAS.

The two RNAs are redundant

Deletion and mutation of the roX RNAs has shown that they are redundant; only when both are disrupted is male survival impacted (Meller et al., 2000).
and Rattner, 2002). Despite their redundancy and their similarities in terms of expression, interaction partners and localisation in immunostaining experiments, the two roX RNAs are quite distinct in sequence. There is only a very small, 30 nucleotide long region that is similar between roX 1 and roX2 (Franke and Baker, 1999). This region is located at the 3' end of both RNAs.

Systematic deletions of regions of the long roX1 transcript have shown that not all of it is necessary for function. In fact, only the deletion of the 3’ end of the RNA failed to rescue flies of a roX1–roX2 background and disrupted its proper localisation inside the nucleus (Stuckenholz et al., 2003). In roX2, which has several alternative transcripts, all of them include the exon where the similar region between the two RNAs region is located (including, of course, the prevalent form) (Park et al., 2005).

When the genomes of 12 Drosophila species were made available, a comparative study identified motifs conserved across the different species in predicted stem-loop regions in the 3’ end of both RNAs (Park et al., 2007). A follow-up study showed that these motifs, now named "roX boxes", are redundant, necessary for H4K16 acetylation and may account for the functional redundancy of roX1 and roX2 (Park et al., 2008). These results together indicate the importance of the motifs in the 3’ region of roX1 and roX2 transcripts.

**Mle is necessary for the proper functioning of the roX RNAs**

Even with the knowledge that the roX RNAs are necessary for the proper functioning of dosage compensation in Drosophila, their exact molecular function and the way they interact with the other members of the complex still elude us. For example, we do not know if it only interacts directly with some of the members of the complex, as they all precipitate together in most conditions. One exception is in conditions of elevated salt concentration, where Mle does not co-precipitate with the rest of the complex (Akhtar et al., 2000). However, this RNA helicase was shown to be necessary for proper roX localisation to the X chromosome and suggested to participate in the early packaging of the RNA into the complex (Meller et al., 2000). Mle is thus a necessary part of the dosage compensation mechanism, even if not an integral part of the DCC in all conditions.

Another member of the complex, Msl2 has not been annotated with a RNA-binding domain, but has been show to interact with RNA in vitro, even if with a lower affinity than it binds to cDNA (Fauth et al., 2010). In addition, roX2 incorporation in the complex was shown to be blocked in a Msl1 mutant that cannot bind Msl2 (Hallacli et al., 2012).

Which proteins directly bind to the roX RNA and in what fashion they do so has not been yet investigated. The evidence that Mle is necessary for packaging the roX RNAs into the DCC and that Msl2 also
binds DNA makes those the most promising candidates for interaction with the roX RNAs.

3.1.2 iCLIP: probing protein-RNA interactions at the nucleotide level

In genomics, it is routine to identify the DNA binding sites for proteins such as transcription factors using ChIP-seq (see chapter 1). A similar technique has been developed that is analogous to ChIP-seq, but that identifies the interactions between a protein of interest and RNA: UV Cross-Linking and Immuno-Precipitation (CLIP, Ule et al., 2005). In CLIP, proteins are cross-linked to RNA using ultraviolet radiation and an antibody for the protein of interest is used to purify the protein-RNA complexes. The RNA fragments are then converted to DNA using a reverse transcriptase. These are sequenced and mapped to the reference genome of the organism to identify the binding sites. There are several variations of CLIP, including Photoactivatable-Ribonucleoside-Enhanced CLIP (PAR-CLIP, Hafner et al., 2010), where the cells have been cultured with photoactivable nucleosides to increase the efficiency of UV cross-linking, and, Individual-nucleotide resolution CLIP (iCLIP, König et al., 2010) which has improved the capability of CLIP to probe interactions at nucleotide resolution over the original CLIP and PAR-CLIP.

To achieve this precise resolution, iCLIP digests most of the cross-linked protein, leaving a small polypeptide at the cross-linked nucleotide. This peptide then blocks the reverse transcriptase, so that the cDNA fragment truncates at the cross-linked nucleotide. iCLIP also includes a random barcode in the sequencing adaptors that allows for the distinction between individual binding events, when several reads map to the exact same position. iCLIP allows us to map protein-RNA interactions both on a genomic scale and at a single-nucleotide resolution.

3.1.3 Objective

We set out to identify the precise locations where Mle and Msl2 bind the roX RNAs. Using iCLIP we created high-resolution datasets of RNA binding for these two proteins in a male-derived D. melanogaster cell line. We also identified Mle’s interaction with DNA via ChIP-seq. Further, we characterised the secondary structure of the roX RNAs and studied how this relates to the interaction with the rest of the DCC. Knowing where the DCC proteins interact with the roX RNAs, we can identify why the two are redundant and why the roX boxes are essential for their function.
3.2 Results

3.2.1 Mle binds DNA mainly in the X chromosome

To characterise Mle’s binding to DNA, we performed a ChIP-Seq experiment using an antibody against Mle in a *D. melanogaster* Schneider 2 (S2) cell line (see chapter 1 for a description of the technique). We produced a single IP sample and a corresponding control from input DNA.

Upon completion of the sequencing of the immunoprecipitated sample and the control input I mapped the reads to the reference *Drosophila* genome using Bowtie (Langmead et al., 2009) and identified the areas where Mle binds the genome (see chapter 2 for a detailed description of the method and a list of publications where it has been used).

The number of sequenced reads for the ChIP-seq experiment can be found in Table 8. Sequencing of the IP sample yielded almost 28.5 million reads, of which over 17 million (60.19%) mapped to the *Drosophila* genome. As for the control input, we sequenced close to 31 million reads, with more than 29.5 million (95.86%) of them mapping to the reference genome. There are 843 contiguous regions bound by Mle, with an mean width of 370 bp. In total, around 312 kb of the genome were classified as being bound by Mle, 305 kb of these in the X chromosome.

<table>
<thead>
<tr>
<th>Sample</th>
<th># reads sequenced</th>
<th># reads mapped</th>
</tr>
</thead>
<tbody>
<tr>
<td>IP</td>
<td>28,458,552</td>
<td>17,129,791</td>
</tr>
<tr>
<td>Input</td>
<td>30,856,427</td>
<td>29,579,043</td>
</tr>
</tbody>
</table>

Table 8: Reads sequenced and mapped for ChIP-seq

A manual overview of the binding signal over the genome suggested that Mle had sharper, more defined binding regions compared to other members of the DCC, such as Mof, which on the X chromosome can cover the entire gene body. In addition, the peaks seemed to localise mainly to the X chromosome, overlap with Mof and with previously defined High-Affinity Sites (HAS (Alekseyenko et al., 2008; Straub et al., 2008). An area of the genome where this overlap between Mle, Mof and HAS occurs can be seen in Figure 3.14a, where a region of extensive Mof binding (in yellow) overlaps several genes and coincides with a much sharper peak of Mle (in green). The Mle peak overlaps neatly with a HAS, marked in red.

To confirm these ad-hoc observations, I investigated the location of the genes that overlapped with regions significantly bound by Mle. As can be seen in Figure 3.14b, the majority (90%) of the 489 genes bound by Mle are located in the X chromosome. This agrees with the microscopy images obtained by Mle immunostaining where Mle paints the X chromosome extensively.
Figure 3.14:  (a) An example of Mle binding in the X chromosome. The red boxes represent the location of HAS, the green track represents the binding of Mle and the orange one the binding of Mof. RefSeq gene annotations for both strands are below.  (b) Genes bound by Mle in a ChIP-Seq experiment. Most genes bound are located in the X chromosome.  (c) Many genes bound by Mle in the X chromosome also overlap a HAS. The top track shows all genes in the X chromosome in grey and Mle-bound genes in green. The middle track shows all X-linked genes in grey and the location of the HAS in red. The bottom track shows the location of Mle-bound genes that overlap a HAS, with the location of roX1 and roX2 marked by the blue arrows.
Some of the DCC components, like Mof and Msl1, have a very distinct binding profile along the gene, as seen in chapter 2 and exemplified for Mof in Figure 3.14a. To see whether this is the case for Mle, I aligned all the transcription start sites (TSSs) and transcription termination sites (TTSs) from *D. melanogaster* genes, as defined by their Ensembl gene start and end coordinates [Flicek et al., 2012], and plotted the average signal of the bins that are bound around them to create a metagene representation of Mle binding. The metagenes for the entire genome, for the X-linked genes and for the autosomal genes are shown in Figure 3.15a. It is clear that there is no enrichment at the promoters or in the gene body, as it is the case for Mof and Msl1. In the metagene composed from just the genes bound by Mle (Figure 3.15b), there is a increase in binding at the TTS as well as the gene body. As we will see later in this chapter, Mle interacts mainly with the roX RNAs. The higher binding of Mle on the 3’ end of the gene is in agreement with previous findings that roX2 localises to the TTS of X-linked genes [Chu et al., 2011].

The fact that Mle seemed to overlap with High-Affinity Sites (HAS) in the genome browser evaluation led me to consider whether this happened on the whole X chromosome. Two studies have identified HAS in a genome-wide fashion, using approaches that determine via ChIP-seq or ChIP-on-chip the locations where individual members of the DCC still bind to the X chromosome in conditions where one or more of the DCC components is missing. Alekseyenko et al. (2008) identified 131 sites bound by Msl2 in male flies with a msl3 background. These sites contained a GA-rich motif and had an average width of 1641 bp. Straub et al. (2008), on the other hand, identified 150 sites that show Msl1 and Msl2 binding in S2 cells, in which the levels of Mof, Msl3 and
Mle had been reduced by RNA interference (RNAi) or where Ms1 and Ms2 binding to DNA is still detected despite lower formaldehyde cross-linking. The sites found by Straub et al. (2008) are narrower, with an average width of 1055 bp. The HAS found by both studies overlap to a great extent: the union of the 131 sites from one study and the 150 from the other results in 189 distinct HAS.

As expected, the overlap between Mle-bound genes and the 189 HAS is quite high. Almost all of the HAS (170, 89.9% of the total) are contained inside genes that are also bound by Mle. Conversely, 204 (41.7%) of Mle-bound genes in the X chromosome overlap a HAS. The ChIP-seq experiment shows that the DNA-binding function of Mle acts mainly on the X chromosome and is closely related to HAS.

3.2.2 Mle RNAi lowers the levels of all DCC components

As described in the introduction, Mle is the only member of the dosage compensation complex that does not co-precipitate together with the other complex members under high-ionic conditions. Despite this weaker association, its presence has an impact on the other members of the DCC. This is apparent in an RNAi experiment we carried out that depleted the levels of Mle in the cell. As shown in Figure 3.16a, in the cells where Mle RNAi was carried out the level of all other protein members of the complex decreased.

This experiment was carried out on a clone8 cell line. Unlike S2, these cells express both roX1 and roX2 and so allow us to see the behaviour of both RNAs. The relative levels of roX1 and roX2 decreased drastically in the Mle knockdown (Figure 3.16b).

These results confirm that the DCC needs the presence of Mle, even though it shows different binding profiles on the DNA and complexes without it have been isolated.

3.2.3 Mle binds mainly roX RNAs

In addition to being a DNA helicase, Mle has RNA helicase activity. After having evaluated the binding of Mle to DNA and showing that its depletion leads to the loss of the other complex members, we decided to explore the binding of this protein to RNA. To do this, we performed an iCLIP experiment using an anti-Mle antibody. This experiment was performed in three biological replicates in clone8 cells.

As described in subsection 3.1.2 of this chapter, iCLIP’s random barcoding strategy allows for a better discrimination between genuine binding events to the same nucleotide and mere PCR amplification artefacts. We included an extra, non-random, barcode to allow for the sequencing of the three replicates in a single multiplexing lane. After separating the reads according to their replicate barcode, replicate 1 had 50,307,338 reads, replicate 2 had 9,434,882 and replicate 3
Figure 3.16: DCC members are depleted upon Mle RNAi. (a) Western blot analysis of Mle and GFP RNAi experiments. All protein members of the DCC show lower levels when Mle is depleted. Tubulin and NXF1 were used as loading controls. (b) qPCR analysis after reverse transcription of roX1, roX2 and PGK from total RNA extracted from cells subjected to Mle and GFP RNAi. Both DCC RNAs are depleted in the absence of Mle. PGK was used as a control.

had 34,728,320. I mapped the reads to the reference \textit{D. melanogaster} genome using BWA \cite{Li2009}. Between 11\% and 18\% of each replicate’s reads correctly mapped to the genome. To discriminate between spurious and significant binding, I used a technique similar to that of König \textit{et al.} \cite{Konig2010}: for every region (exons, introns) of a transcript, I created a thousand permutations of the actual observed binding events after clustering those events happening within 30 nucleotides, and calculate the probability that the observed number of events would come up by chance. This allowed me to choose a cut-off that results in a False Discovery Rate of 0.05. This procedure was applied for every replicate separately. After, I kept those nucleotides bound by two out of three replicates. To these nucleotides I assigned the conservative lower number of individual binding events (as determined by the number of unique random barcodes) of the three replicates. This left a set of 6525 individual binding events at 956 distinct nucleotides.

With data describing the locations where Mle binds to the RNA molecules, I could investigate where their coding genes are located. The locations of the DNA origin of the transcripts bound by Mle can be seen in Figure 3.17a. In this figure the locations of roX1 and roX2 have been highlighted in red and blue, respectively. It becomes clear that these two transcripts show the strongest binding by Mle. However, in contrast to its DNA-binding behaviour, Mle binds RNAs from all \textit{Drosophila} chromosomes. In fact, out of the 193 genes whose transcripts are bound by Mle, only 31 are located in the X chromosome.
Looking deeper into the location of the nucleotides bound by Mle, we can see in Figure 3.17b that the majority (72.7%) of Mle-bound nucleotides belong to the DCC RNAs, roX1 and roX2. After them, the most bound nucleotides are those belonging to protein-coding genes, with a total of 249 bound nucleotides. Inside these, most (172) belong to intronic regions.

3.2.4 Msl2’s main targets are also the roX RNAs

We followed up the analysis of Mle binding to RNA with another iCLIP experiment, this time with an anti-Msl2 antibody. This experiment followed the same experimental design and data-processing as the Mle iCLIP and was also done in three biological replicates. The first replicate had 32,445,515 reads, the second 44,093,275 and the third 1,358,360. For the different replicates, 12.25%, 8.89% and 49.31% mapped to the reference genome.

Overall, Msl2 binds markedly fewer RNAs than Mle, as one would expect due to its lack of a clear RNA-binding domain. A total of 539 nucleotides are bound by Msl2, distributed among the transcripts of only 45 different genes. Of those 45, 31 come from autosomes and 14 come from the X chromosome. Unsurprisingly, roX1 and roX2 are the transcripts most abundantly bound by this protein (Figure 3.18a), with 90.7% of the bound nucleotides belonging to these two RNAs. As with
Mle, most of the remaining bound nucleotides are located in the introns of protein-coding genes (Figure 3.18b).

3.2.5 roX1 and roX2 show defined regions bound by Mle and Msl2

As seen in the previous sections, both Mle and Msl2 show abundant binding for both the roX1 and roX2 transcripts. I looked into more detail at where exactly in those transcripts the proteins are binding.

At 3.7 kb, roX1 is by far the longer of the two roX RNAs. Mle and Msl2, however, do not indiscriminately bind the whole transcript. Instead, both of them occupy three distinct regions (Figure 3.19a).

The 3’ distal domain of binding by the two proteins coincides with the area of roX1 that is conserved between roX1 and roX2 across the *Drosophila* genus. It is in this domain that the previously described roX box motifs are located. The lower panel of Figure 3.19a shows just this third domain and in it the location of the three motifs is marked in grey, highlighting the fact that all three have clear peaks of Mle binding and two show peaks for Msl2.

While smaller than roX1, roX2 is also not bound in its entirety. Instead, it shows a single domain of distinct binding by both Mle and Msl2 (Figures 3.19b), not unlike the distinct regions bound in roX1. This roX2 region bound by the DCC proteins includes the previously described elements that are conserved between the two roX RNAs. Also like in roX 1, the roX box motifs (again marked in grey in the figure) all contain nucleotides that are highly occupied by Mle.

It becomes clear when comparing the Mle and Msl2 profiles from Figure 3.19b that in this RNA Msl2 shows a much more diffuse pattern...
Figure 3.19: (a) Mle and Msl2 binding to the roX1 transcript. Despite the length of the RNA, three clear and distinct regions of binding are present. The lower panel shows the region of binding where the roX boxes are located. The precise location of the roX box motifs is highlighted in grey. (b) Mle and Msl2 binding to the roX2 transcript. In roX2 both proteins also bind a clear and defined area inside the RNA molecule, which is shown in more detail in the lower panel, with the location of the roX box motifs highlighted in grey.
of binding than Mle. However, it should be stressed that both proteins bind the same region of roX2.

3.2.6 Mle binds RNA stems

After establishing that Mle and, to a lesser extent Msl2, bind the clearly-defined areas in the roX RNAs and, more specifically, the previously described roX box motifs, I investigated the binding in terms of the secondary structure of the RNA. Using RNAfold to predict the secondary structure of roX2, I could determine whether the nucleotides bound by Mle are in single-stranded or double-stranded areas.

![Figure 3.20: Binding of Mle to double- and single-stranded nucleotides of roX2, respectively. The secondary structure of the RNAs was calculated using RNAfold. In both RNAs the nucleotides with higher binding scores are in double-stranded stems.](image)

As is clear in Figure 3.20, the nucleotides that have high binding score, as measured by the number of individual binding events observed, are in regions predicted to be double-stranded. This, together with the small conserved regions among these two redundant RNAs suggests that it is the secondary structure provided by the roX box motifs that is necessary for the correct incorporation of either of the roX RNAs into the complex by Mle.

3.2.7 SHAPE reveals roX structures

Having established that the binding of Mle to both roX1 and roX2 is linked to the well defined roX box motifs that appear in predicted stem-loop structures, we assayed the structure of the two RNAs more deeply. The technique used was selective 2’-hydroxyl acylation analyzed by primer extension (SHAPE). This technique relies on the higher reactivity of the ribose 2’-hydroxyl group to hydroxyl-selective electrophiles of single-stranded RNA bases (Wilkinson et al., 2006) and allows for
the determination of whether the individual bases of short RNAs are single- or double-stranded.

While assaying the whole length of roX1 and roX2 is not feasible due to limitations of SHAPE, it is possible to assay the stretches of both RNAs that are bound by Mle. We therefore used SHAPE to determine the secondary structure of the domains of roX1 and roX2 that have the top peaks of Mle binding (shown in the bottom panels of Figure 3.19a and b, respectively).

The secondary structure of the 3’ domain of roX1, has three areas that are double-stranded (Figure 3.21a). The first of these, roX1-helix1 (marked R1H1 in Figure 3.21a), contains a motif very similar to the roX boxes, which we termed roX box-like (RBL). The next one, rox1-helix2 (R1H2), includes the first of the roX boxes pairing with an inverted version of the motif. Finally, there is a small stem-loop (P2) which does not include any of the roX-related motifs.

The structure of roX2 gleaned from SHAPE also show that the secondary structure coincides with roX motifs. The first three identified helixes, roX2-helixes 1 to 3 (marked as R2H1-3 in Figure 3.21b), all have the RBL motif seen in R1H1. Two small stem-loops (P3 and P4) do not show any motif. These are followed by another three helix regions (R2H4-5) which overlap the three roX box motifs.

The fact that several of the secondary structures highlighted by SHAPE overlap with roX box or RBL motifs raises the question of whether they coincide with the peaks of Mle and Msl2 binding. By overlapping the binding profiles with the locations of the helixes we can see that is indeed the case (Figures 3.21c and 3.21d). Most of the double-stranded regions have peaks of Mle and, to a lesser extent, Msl2 binding.

Taken together, these results highlight the importance of the conserved region between the two roX RNAs in terms of their binding to the DCC proteins, and suggest that the conservation seems to be related to their secondary structures.

3.2.8 roX2 stem structures are essential and redundant for male viability

Given that the roX RNAs are bound by the DCC proteins in regions that form double-stranded secondary structures, we hypothesised that disrupting these structures would prevent binding and so destabilise the complex and, consequently, dosage compensation in the male flies. To test this hypothesis we created a set of twelve transgenic flies expressing different mutated roX2 variants in a background that has no endogenous roX1 and roX2 expression. The mutations were chosen to disrupt several of the stem structures found in roX2 and were done in three groups: four transgenic lines had mutations that disrupt helixes that contain the RBL motif (R2H1-3), four others had mutations that
Figure 3.21: SHAPE reveals the structure of roX RNAs. (a) The structure of the third domain of roX1 includes three double-stranded regions. Nucleotides are coloured according to their reactivity, with those more likely to be single-stranded in red and those more likely to be double-stranded in grey. (b) The roX2 secondary structure has eight distinct double-stranded regions. (c) The double stranded regions of roX1 and roX2, respectively, overlap with the main peaks of Mle
disrupt helixes that contain the roX box motif (R2H4-6), and the final four had mutations that disrupt helixes belonging to both the R2H1-3 and R2H4-6 regions. We then measured how effective the different transgenic roX2 lines were in rescuing males in this lethal background compared to the wild type roX2.

The results are summarised in Figure 3.22. The mutations did not greatly diminish the stability of the RNA (see right bar charts). Nor did they have a strong effect on male viability when the mutations were only on RBL stems or only on roX box stems (Figures 3.22a and b, respectively). The exception was in the cases where the two 3’ distal structures were disrupted (R2H5 and R2H6, disrupted in the transgenic lines A0B4 and A0B5). It was in these structures that the highest binding peaks for Mle were observed.

More dramatically, the viability is close to zero in all transgenic lines where the mutations disrupt helixes from both the RBL group of helixes in addition to R2H5 and R3H6 (Figure 3.22c). These results show that the double-stranded structures of roX2 that are bound by Mle are essential for proper male development and establish the redundant nature of the conserved binding sites in the RNA.

3.2.9 **Each DCC only has one of the roX RNAs**

While previous work has shown that the roX RNAs are redundant, the question of whether both of them are present in the same dosage compensation complex has not been settled. Using Chromatin isolation by RNA purification (ChIRP) we can establish that each complex only has one species of the roX RNAs. This technique, developed by the Chang laboratory, is similar to ChIP, but instead of antibodies it uses single-stranded oligonucleotides which are complementary to the RNAs that is being probed for (Chu et al., 2011).

For our experiment, we used probes against roX1, roX2 and, as a negative control, LacZ. After isolation, the recovered levels of roX1, roX2 and, as another negative control, GAPDH were measured. If each MSL complex in the nucleus includes a copy of roX1 and another of roX2, we expect to recover both variants of the RNAs when using either of the probes. If, on the other hand, each complex only has roX1 or roX2, we should only recover the RNA that is probed for.

As can be seen in Figure 3.23, the results show very clearly that a probe for roX1 recovers only complexes with roX1 and a probe for roX2 recovers only complexes with roX2. This result strongly suggests that, in addition to being genetically redundant, roX1 and roX2 are interchangeable in the individual dosage compensation complexes present in the nucleus.
Figure 3.22: Male viability is diminished when roX secondary structures are disrupted. The disrupted double-stranded regions are marked in orange on the left panels. The middle panels show the rescue rate for each transgenic line and the rightmost panels show the stability of the mutated RNAs. roX2 with mutations in the RBL structures rescue the male flies in levels comparable with the WT RNA (A0B0). When the two 3’ distal structures (R2H5 and R2H6) are mutated, the recovery rate is greatly decreased compared to the WT RNA. If, in addition to R2H5 and R2H6 the RBL structures are also disrupted, roX2 stops being able to recover males from the deleterious background.
3.3 Discussion

We have established that Mle and Msl2 interact directly with roX1 and roX2 in vivo. The binding does not occur indiscriminately throughout the transcripts, but instead is focused on three distinct areas of roX1 and one area of roX2. The 3’-terminal area of roX1 and the bound area of roX2 correspond to the regions where the roX box elements known to be essential for proper activity are located. Mle has well-defined peaks of interaction with the roX RNAs and the bigger of these overlap both with the functional elements of the transcripts and with nucleotides predicted to form double-stranded regions. Msl2 has a more diffuse binding pattern, but it clearly overlaps with Mle. Since other lncRNAs have been shown to act as mediators that tether distinct protein complexes, it is not hard to imagine the roX RNAs as mediators that tether the DCC members together.

Having determined the structure of the two RNAs, we can also show that the conserved motifs between the RNAs appear in double-stranded regions and that the higher peaks of Mle binding appear in these double-stranded regions. This suggests that the mechanism by which Mle interacts with the roX RNAs recognises not only their sequence but also their conformation. This is confirmed when disruption of the secondary structure of the stems leads to a almost total loss of viability in males. Given the link between RNA secondary structure, Mle binding and male viability, it seems safe to postulate that the conservation across the two RNAs and Drosophila species of these structures is maintained to ensure male viability.
In contrast to previous viability studies, which carried out deletions, we establish that the simple disruption of stem loops in roX2 is responsible for loss of male viability, when the right combination of binding sites is disrupted. This highlights the redundancy present inside roX2. The corresponding region of roX1 where the roX boxes are located shows a similar pattern of structure and Mle binding, suggesting that the same relationship between sequence, RNA structure and Mle binding will hold for this RNA as well.

Intriguingly, roX1 has two other domains of Mle binding along its length. In previous studies the deletion of just the 3’ end of roX1 was enough to drastically reduce male viability, so these might not function as redundant functional regions. Given that the levels of Mle binding at these regions is lower, it might be the case that they do not show enough activity for proper Mle recruitment. Nevertheless, it is clear from our results that in order to ensure male viability the roX RNAs provide several layers of redundancy: the two different roX RNAs and the different binding sites inside each transcript.
RNA POLYMERASE II HAS DISTINCT MODES OF OCCUPANCY IN A HIGHER EUKARYOTE

The work presented in this chapter uses data from a variety of publicly available sources. Florence Cavalli performed the initial processing and provided the files with the log$_2$ fold change of IP over input of the RNA Polymerase II ChIP-seq experiments from Conrad et al. (2012b). For other data sources the publicly available processed files were used, except where noted.

4.1 INTRODUCTION

4.1.1 Transcription as an interplay of multiple components

As described in chapter 1, eukaryotic transcription is a complex process that involves an interplay of promoter sequence motifs, basal transcription factors, sequence-specific transcription factors and the RNA Polymerase II (RNA Pol II) machinery itself. After Pol II has been released from the promoter its activity can still be regulated. The polymerase can be stopped shortly after the transcription start site (TSS), in an event called stalling that helps regulate developmental genes (Zeitlinger et al., 2007). Even in cases where there is no stalling, the various steps of transcription are regulated. After assembly, the release of Pol II from the promoter only happens after CDK7 phosphorylates serines 5 and 7 (Ser5 and Ser7, respectively) of Pol II’s C-Terminal Domain (CTD). The phosphorylation of the CTD also changes during the other steps of the transcription cycle. During elongation, Ser5 and Ser7 are dephosphorylated, and Ser2 is phosphorylated instead, while during termination Ser2 itself is dephosphorylated to return Pol II back to its unphosphorylated state (for a review of CTD phosphorylation during transcription see Hsin and Manley, 2012).

4.1.2 Drosophila melanogaster is a well-studied system

Recent years have seen an explosion in genome-wide studies of chromatin and transcription-related factors, aided by new techniques and projects like the Drosophila model organism Encyclopedia of DNA Elements (modENCODE). This project, which had its main data release in December 2010 (modENCODE Consortium et al., 2010), has followed up with detailed studies of the transcriptome, cis-regulatory elements and chromatin landscape of D. melanogaster.
Using a mixture of microarray and extensive RNA-seq expression profiling, Graveley et al. (2010) provided an overview of the transcriptome of D. melanogaster during development. Twenty-seven developmental stages were covered, from the 2 hour-old embryo to adult flies 30 days after eclosion. The main finding from this study was the extent to which D. melanogaster genes have alternative transcripts: a three-fold increase in the number of alternative splicing events compared with previous results.

Nègre et al. (2011) characterised the interactions between transcription factors (TFs) and DNA and the co-localisation of TF binding sites. Using a combination of ChIP-chip and ChIP-seq datasets, they mapped the binding sites of 38 TFs at several developmental stages. They found a great deal of co-binding by the studied TFs, with 38.3% of binding sites being bound by more than two factors. Remarkably, a small number of sites was bound by more than eight factors and was thus characterised as high occupancy targets (HOTs). These kind of regions have previously been found in Caenorhabditis elegans. Their function is unclear. Some studies have reported that despite the vast number of TFs binding in HOT areas, they are not enriched for the DNA motifs of the TFs that bind there (Moorman et al., 2006; modENCODE Consortium et al., 2010). Distinct results were found for the genes influenced by HOT regions: in C. elegans genes from these regions were expressed in most cell types (Gerstein et al., 2010), while Kvon et al. (2012) found that in D. melanogaster these regions act as enhancers with distinct spatial and temporal patterns, in a behaviour similar to that of developmental genes. In both cases, however, genes near HOT regions are more likely to be highly expressed. One hypothesis is that these regions of extensive TF co-binding are a result of high TF concentrations in the nucleus, leading to their binding these areas of open chromatin (MacArthur et al., 2009).

The use of whole embryos in some of these genome-wide studies leads to a spatial limitation in the results that they produce. By taking material from all the tissues and cell types present in the embryo, differences in TF binding or expression patterns between these cell types are not visible. Not only is it not possible to distinguish tissue-specific binding sites from ubiquitous ones, but those binding events that are only present in a small number of cells might not produce enough signal to be visible when the chromatin from those cells is mixed with that from all the other cells in the embryo. A possible approach to map binding of proteins to DNA in manner that is both genome-wide and tissue-specific is to batch isolate tissue-specific chromatin for immunoprecipitation (BiTS-ChIP, Bonn et al., 2012). In this method, a tagged nuclear protein is expressed via a transgene in the cell type or tissue of interest. After cross-linking of the proteins and DNA, the nuclei are sorted by flow assisted cell sorting (FACS) to obtain pure populations
of cross-linked nuclei expressing the tagged protein, which are then immunoprecipitated and sequenced.

On a more restricted set of conditions (S2 and BG3 D. melanogaster cell lines), (modENCODE Consortium et al. 2010; Kharchenko et al. 2010) mapped the locations of 18 different histone modifications. Using a Hidden Markov Model (HMM), they defined nine different chromatin states that resulted from combinations of the original 18 histone modifications. These states in turn corresponded to different functional areas of the genome. One of the states, for example, identified promoters via the presence of di- and trimethylation of the lysine 4 and the acetylation of lysine 9 of the histone H3 (H3K4me2, H3K4me3 and H3K9ac, respectively). Another state identified areas under transcriptional elongation via the presence of trimethylation of lysine 36 of histone H3 and H2B ubiquitination. The particular behaviour of H4K16ac in the X chromosome led to yet another state that marks this chromosome extensively. The authors also defined a more complex map of thirty chromatin states.

In another example of using a variety of proteins to describe overarching chromatin states, Filion et al. (2010) used DamID to identify the binding sites of an extensive catalogue of chromatin proteins and combined that data into “colours”, representing combinations of proteins that are associated with different modes of expression or repression of transcriptional activity.

Other studies outside modENCODE have focused on specific systems and processes, also on a genome-wide scale. Mavrich et al. (2008) mapped nucleosome positioning in the fly and identified several differences to the well-characterised nucleosome architecture of Saccharomyces cerevisiae. Unlike yeast, in fly the first nucleosome after a promoter is present after the TSS and not on top of it. Also unlike yeast, the preceding nucleosome does not comprise the H2A.Z variant. Gilchrist et al. (2010) and Nechaev et al. (2010) studied Pol II stalling across the genome from the perspective of nucleosome positioning and short RNA synthesis, respectively, and found that stalling is both widespread and present at highly regulated genes. One possible role for the pausing of Pol II is to permit the synchronisation of expression of developmental genes between the different cells of an organism. Lagha et al. (2013) demonstrated that diminishing pausing in snail promoters results in a less coordinated expression of that gene and disturbs mesoderm invagination in the developing fly embryo.

The particular behaviour of Pol II in X-linked dosage compensation was the focus of two conflicting studies. The first used Global Run-On Sequencing (GRO-seq) which measures nascent RNA to identify the locations of engaged Pol II. The authors report increased levels of elongation on the gene bodies of X-linked genes compared to autosomal genes in S2 cells, which have a male phenotype (Larschan et al. 2011). A second study, on the other hand, produced ChiP-seq datasets of Pol
II in both male and female samples and reported an increase of Pol II occupancies across the entire X-linked genes in males, including the promoter, compared to the autosomal genes. This difference was not observed in the female samples. Furthermore, they did not observe an increase in elongation, concluding that dosage compensation in fly is brought on by an increased recruitment of Pol II at the promoter of X-linked genes (Conrad et al., 2012b; Vaquerizas et al., 2013).

4.1.3 An unbiased study from RNA Polymerase II’s perspective

The current landscape of genomics research shows two strong trends: the generation of vast and diverse datasets and the need for computational tools to analyse, compare, contrast and draw conclusions from those datasets. Given the open-data policy of biology, where datasets are deposited on public repositories or made available directly through their consortia websites upon publication, one can use publicly available data and analyse them in new manners to identify previously unknown biology.

With this objective using I have analysed the behaviour of RNA Polymerase II publicly available datasets and have found distinct modes of action of this complex. The occupancy profile of Pol II changes in shape and location across the different modes. These in turn correspond to differences at the promoter sequence, regulatory TF binding and functional annotation levels.

4.2 Results

4.2.1 Analysis of Pol II datasets

The RNA Pol II ChIP-seq datasets used in this analysis were published in Conrad et al. (2012b) and were obtained using an antibody against the Rpb3 subunit of Pol II in male and female third instar larvae salivary glands and in S2 cells.

The meta profiles of Pol II for all promoters in D. melanogaster can be seen in Figure 4.24. Pol II signal rises to a peak around 25 bp upstream of the TSS, then descends into the gene body. When manually surveying the Pol II signal in the genome, there was the suggestion that it doesn’t behave in a similar way across all TSSs. For this reason I tried to identify the different shapes that the peak of Pol II takes across different promoters.

One way to describe the profile of the peak at the TSS is to use the derivative of the signal. A widely used mathematical concept, the derivative is a measure of the slope of the curve of a function. It takes positive values when the curve has an upwards trajectory, is zero when it is flat and has negative values when the curve is descending. As it describes only the slope of a curve, the derivative is the same for any
parallel lines. This means that when describing the occupancy of Pol II at a promoter, two genes that have profiles with the same slopes and peak locations but different overall occupancy levels, will have the same derivatives. This property makes the derivative a good way to capture differences in Pol II peak location and sharpness, regardless of the overall intensity level.

To see whether there were indeed different Pol II binding profiles, I first created a set of high-confidence TSSs. I downloaded the TSS for all transcripts in the Ensembl database (Flicek et al., 2012). A recent dataset of short capped RNAs from D. melanogaster samples (Nechaev et al., 2010) was used to update the TSS positions with experimental data. To use the data from both male and female ChIP-seq datasets without the confounding influence of differential behaviour between the sexes, two measures were taken. First, given that in D. melanogaster the X chromosome is up-regulated as a whole and shows different Pol II behaviour, all transcripts from this chromosome were excluded. Second, transcripts that are differentially expressed between the male and female samples were excluded. From these constraints resulted a total of 15513 unique TSSs which were used for the analysis.

For the resulting set of 15513 TSSs, I extracted the signal intensity values from 300 bp upstream to 700 bp downstream of the TSS in both male and female samples. I then converted each of these promoter vectors to their derivative. Using k-means clustering on the resulting matrix, I identified five clusters with distinct Pol II profiles. The clusters found were similar when using just the male or female samples or when changing the number of clusters.

4.2.2 Pol II shows distinct profiles across TSSs

The heatmaps of Pol II values and derivatives, sorted by the clustering results, are shown in Figures 4.25a and 4.25b, respectively, and the average profile for each cluster is shown in Figure 4.25c. Some facts about the clusters are obvious from the figures. The first is that they differ in the number of transcripts that they include. Cluster 1, with 7712 TSSs, is
RNA POL II HAS DISTINCT MODES OF OCCUPANCY IN FLY

Figure 4.25:  
(a) Heatmap of RNA Pol II signal around the TSSs of *D. melanogaster*. 
(b) Derivative of the signal of RNA Pol II around the TSSs. The differences in shape of the profiles of Pol II are brought to evidence by the derivative and allow us to cluster the TSSs in five distinct groups. 
(c) The average profile for each of the clusters illustrates the different shapes that the peak of Pol II signal takes around the promoters.
the most abundant, followed by cluster 2 (2753 TSSs), cluster 3 (1796 TSSs) and finally cluster 4 (1233 TSSs). The average profiles for each of the clusters make evident their different profile shapes: cluster 1 has a flat profile, without any apparent peaks; cluster 2 shows a broad peak around the TSS; cluster 3 shows a more pronounced peak in the same location as cluster 2; cluster 4’s peak is over 300 bp upstream of the TSS; and cluster 5 also has a peak that has been shifted upstream, but only to 125 bp upstream of the TSS. These differences might not be as easy to discern on the heatmap of the original values (Figure 4.25a), but they are there nonetheless, in the left shift of high (blue) values of clusters 4 and 5, and the more pronounced increase of values around the TSS of cluster 3 compared to cluster 2.

![Figure 4.26](image)

Figure 4.26: The clusters found are robust and maintain their shape when the transcripts are divided between those that have another transcript within 500 bp upstream or not (a), or between those for which the previous transcript is in same or the opposing direction (b).

To check if proximal genes are affecting the Pol II profiles in clusters 4 and 5 (in which peaks are shifted upstream) I plotted the average profiles for the TSSs that have another gene more or less than 500 bp upstream in Figure 4.26a and for those TSSs for which the upstream gene is in the same or opposing direction in Figure 4.26b. In both cases, the profiles do not change shape to a great degree, indicating that the neighbouring promoters are not responsible for the Pol II profile.

**Similar profiles are found in S2 cells**

Having established the presence of distinct Pol II profiles in salivary glands, I looked for their presence in S2 cells. There are two reasons for this. The first is to see where TSSs change profiles between conditions. The second is that there are more datasets for the S2 cell line in
modENCODE that can be directly compared and used to try to explain the presence of the different profiles. For this reason, I used ChIP-seq data for Pol II from S2 cells from Conrad et al. (2012b) to investigate the clusters there. The ChIP-seq processing was done in the same way as for the salivary gland samples.

As the clusters had already been defined in salivary glands, the classification of the transcripts in S2 was done in a different way. Rather than finding de novo clusters, I assigned each TSS in S2 cells to one of the clusters found in salivary glands. To do so, I measured the euclidean distance of the derivative profile of the S2 TSSs to the average derivative profile of each cluster. Each S2 TSS was assigned to the cluster to which it was nearest. The number of TSSs assigned to each cluster in the two samples is given in Table 9. After classifying, I plotted the average profiles of each cluster with the signal from the S2 sample. These are presented in Figure 4.27a, and are remarkably similar to the original profiles from Figure 4.25c.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salivary glands</td>
<td>7712</td>
<td>2753</td>
<td>2019</td>
<td>1233</td>
<td>1796</td>
</tr>
<tr>
<td>S2 cells</td>
<td>8492</td>
<td>3545</td>
<td>2035</td>
<td>970</td>
<td>470</td>
</tr>
</tbody>
</table>

Table 9: Number of TSSs in the salivary gland and S2 samples

Given that I assigned the TSSs to new clusters in the S2 data, it is informative to see if there was any bias in the assignment. Figure 4.27b shows how many TSSs from each salivary gland cluster were assigned to which S2 cluster. Promoters of cluster 1 in the salivary gland samples show almost no change in S2. Clusters 2 and 3 also have the majority of their TSSs from salivary glands retaining that classification in the S2 cell type, but to a smaller degree. Around one quarter of TSS in cluster 2 lose their Pol II peak and are reassigned as belonging to cluster 1, and less than a fifth show a more defined peak and move to cluster 3. Of cluster 3, almost a third of TSSs soften the slope of their peak and move to cluster 2. In cluster 4 close to half of its TSSs are classified as cluster 1 in the S2 sample. There are more pronounced changes in cluster 5, where most TSSs have their Pol II peak shifted downstream and so they are moved to cluster 2 in the S2 sample. The fact that transcripts change clusters between the salivary gland and S2 samples suggests that the clusters are dynamic and can be reflecting changes in the promoters’ expression or chromatin status. In addition, the difference in the groups that see more change between the two samples might indicate that clusters 4 and 5 are more sensitive to those changes.

The different shapes that the Pol II signal takes around the promoters are intriguing. Previous work has shown that the average peak of stalled genes is shifted upstream compared to expressed genes (Conrad et al., 2012b). My analysis is unbiased, in the sense that no information
4.2 Results

(a)

Figure 4.27: (a) S2 clusters closely mirror in shape the clusters found in the salivary gland samples. (b) Change of cluster for the different TSSs. Most TSSs stayed in the same cluster between salivary glands and S2 cell lines. The exceptions are cluster 5 and, to a lesser extent, cluster 4.

about the expression status of the promoters was used to obtain the different clusters. There is the possibility that the clusters are still reflecting expression levels, but also that they represent other underlying characteristics of the different transcripts that make up the different clusters.

4.2.3 Expression analysis

To see if the clusters reflect the expression status of genes, I characterised the expression status of the five different binding profiles of Pol II.

Affymetrix GeneChip datasets were available in triplicates for the same three conditions from which Pol II ChIP-seq was obtained. These were normalised using GCRMA (ger). I classified probes as expressed or non-expressed inside each sample using the Affymetrix MAS5 algorithm. To identify differentially expressed genes I filtered out the probes with an inter-quartile range of less than 0.5, so that only more variable probes would be tested and used limma (Smyth, 2005) on the remaining probes to identify the differentially expressed genes between the male and the female samples that were filtered out when selecting a stringent set of TSSs to cluster.

I used a stringent classification of expression that combines both the microarray and Pol II data to identify expressed or stalled transcripts. For a transcript to be classified as expressed, all the microarray probes had to have positive MAS5 calls, it has to have the Pol II signal at the TSS over the significance threshold determined by the ChIP-seq
processing (log2FC of 1.45 for the male sample and 1.79 for the female) and a positive Pol II log2FC median value along the gene body. Additionally, transcripts were classified as stalled if their microarray probes are marked as negative by MAS5, and they have a negative Pol II log2FC median value in the gene body and Pol II signal on the TSS over the significance thresholds. Finally, a transcript has to be in the same category in both the male and female sample to be so classified.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expressed</td>
<td>242 (3.14%)</td>
<td>375 (13.62%)</td>
<td>395 (19.56%)</td>
</tr>
<tr>
<td>Stalled</td>
<td>176 (2.28%)</td>
<td>206 (7.48%)</td>
<td>147 (7.28%)</td>
</tr>
</tbody>
</table>

Table 10: Number of expressed and stalled transcripts in each cluster and, in parentheses, the same value represented as the percentage of the transcripts of that cluster.

The proportion of each cluster that is classified as expressed or stalled is laid out in Table 10. In agreement with the flat profile of Pol II in cluster 1, it shows the lowest proportion of either expressed or stalled transcripts. Clusters 2 and 3 have the highest proportions of expressed transcripts (13.62% and 19.56%, respectively). By contrast, the shifted clusters 4 and 5 have the highest proportions of stalled transcripts (11.92% and 10.75%, respectively).

Since clusters 4 and 5 show the highest proportion of stalled transcripts, the upstream shift of their Pol II peaks could just reflect this expression status. However, both stalled and expressed genes show similar profiles within each cluster, indicating that the profile shapes are independent of the expression state (Figure 4.28). The shift upstream of clusters 4 and 5 is maintained in expressed genes, as is the Pol II peak on the TSS of cluster 2 and 3 for stalled genes. Thus the shape of each profile is independent of the expression status of the transcripts.

4.2.4 Clusters show different levels of transcriptional fine-tuning over the course of development

Gene expression can vary across an organism’s development and between different cell types. Some genes are essential in all conditions and are expressed accordingly; these are called housekeeping genes. Others are involved in cell type-specific behaviours, such as creating the actin structures of muscle cells, or regulating different developmental steps. This means that housekeeping genes do not need precise regulation, compared with other genes that need to be transcribed at precise times and locations. To test whether the different behaviour of
4.2 RESULTS

Figure 4.28: The profiles of the clusters have similar shapes between expressed and stalled transcripts, the main difference is at the level of the signal.

Pol II in each cluster might reflect different regulatory mechanisms of the promoters, I investigated if the clusters have different expression behaviours during development.

As part of the modENCODE consortium, Graveley et al. (2010) produced a comprehensive RNA-seq dataset for D. melanogaster development. This dataset encompasses 30 different developmental conditions, from the early embryonic stages to adult flies. I utilised this dataset to understand how the different clusters behave during development. As a measure of variability of expression, I used the inter-quartile range of the FPKM values of each transcript over the 30 developmental stages. The box plots summarising each cluster in Figure 4.29 show that transcripts in cluster 3 have higher variability during development than all others. This variability suggests that transcripts from cluster 3 require more fine-tuning in their expression during development that the other clusters.

4.2.5 Promoter types

Metazoan promoters can differ in their underlying DNA sequences. The differing features include basic sequence motifs, or core sequence elements, that are associated with distinct types of promoters that drive the transcription of separate classes of genes, such as housekeeping genes, developmental regulators or tissue-specific genes. By virtue of their different sequences, they are recognised by separate transcription factors which recruit the core transcription machinery.
In *D. melanogaster*, Uwe Ohler’s McPromoter (Ohler, 2006) identified predicted TSSs based on five models of core elements. These five classes of promoters are divided between those that have the Initiator and Downstream Promoter Elements (Inr/DPE), promoters with only the mis-named DNA Replication Element (DRE), those with only the Initiator element (Inr), those with two motifs identified by Ohler (Motif6/Motif1) and finally the canonical promoters with a TATA box and the Initiator element (TATA box/Inr).

The different promoter core elements are associated with different set of genes and TSS architectures. Inr and Inr/DPE promoters are linked with developmental genes (Engström et al., 2007). Motif6/Motif1 and DRE elements, on the other hand, are associated with ubiquitously expressed genes with broad TSSs (Rach et al., 2011).

The analysis of expression in different fly tissues and developmental stages suggested that promoters in cluster 3 have a higher variability of expression during the organism’s growth. If this is an indication that those promoters are involved in the processes that drive development, it is reasonable to assume that those same promoters will have the core motifs associated with developmental genes. To see if this is indeed the case, I assigned the TSSs to one of the promoter models and counted how many TSSs of each cluster belonged to each model. A TSS is assigned to one of the models if it is within 50 bp of the site predicted by McPromoter.

The proportion of each cluster’s TSSs that belong to each promoter class are represented in Table 11. Several trends are visible. Cluster 1, for example, has a low proportion of TSSs classified among all models. The other four clusters, on the other hand, have a clear en-
4.2 Results

Enrichment for DRE and Motif6/Motif1 promoters. In addition, cluster 3 has a higher proportion of Inr and Inr/DPE promoters than all other clusters. The presence of Inr and Inr/DPE promoters in cluster 3 is a strong suggestion that this cluster represents developmental genes, and points towards a functional specificity of the transcripts belonging to this cluster. It is also in agreement with the observation from the previous section that transcripts from cluster 3 show the highest variability across developmental stages.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inr/DPE</td>
<td>2.80%</td>
<td>5.70%</td>
<td>14.21%</td>
<td>0.89%</td>
<td>2.17%</td>
</tr>
<tr>
<td>DRE</td>
<td>2.26%</td>
<td>20.45%</td>
<td>14.86%</td>
<td>12.08%</td>
<td>27.84%</td>
</tr>
<tr>
<td>Inr</td>
<td>2.01%</td>
<td>2.54%</td>
<td>7.97%</td>
<td>1.54%</td>
<td>0.72%</td>
</tr>
<tr>
<td>Motif6/Motif1</td>
<td>1.93%</td>
<td>17.69%</td>
<td>15.70%</td>
<td>10.06%</td>
<td>15.76%</td>
</tr>
<tr>
<td>TATA box/Inr</td>
<td>6.82%</td>
<td>4.18%</td>
<td>6.19%</td>
<td>1.54%</td>
<td>1.34%</td>
</tr>
</tbody>
</table>

Table 11: Proportion of TSSs with a predicted promoter type in its vicinity for each cluster

4.2.6 Clusters show differential TF binding

Having looked into the expression characteristics of the different clusters and their promoter sequences, the question of whether TF binding reflects the differences between the clusters arises. Since both the expression analysis and the comparison of core promoter elements suggest that one of the Pol II profiles in particular is associated with developmental genes, this association is expected to be borne out also in the binding of developmental TFs. To find out, I downloaded the TF binding sites available from modENCODE datasets from third instar larvae, which is the same developmental stage at which the Pol II datasets were obtained. The four TFs available are Caudal, Ecdysone receptor (EcR), Deformed (DFD) and Nejire. All four have been implicated in development and morphogenesis.

To calculate the enrichment of binding in each cluster, I identified the TSSs that have a TF binding site within 400 bp of the TSS. I then determined the enrichment of each cluster for the binding of each TF using equation 1. This calculates how much more likely a TSSs from a specific cluster is to be bound by a TF compared with all other TSSs.

\[
\text{enrichment} = \log_2 \left( \frac{\#\text{TSSsInClusterWithTF}}{\#\text{TSSsInCluster}} \right) \left( \frac{\#\text{TSSsWithTF}}{\#\text{TSSs}} \right)
\]  

(1)

Figure 4.30a shows the enrichment that each cluster has for each TF and for the union of all TF binding sites. For example, the first column shows that cluster 1 is depleted for general TF binding (as evidenced by the blue colour), that cluster 5 is neither enriched nor depleted, and
that out of the three enriched clusters (clusters 2, 3 and 4, in red), cluster 3 shows the highest enrichment for TF binding. Looking at the enrichments of individual TFs in the other columns of Figure 4.30a, cluster 3 shows the highest enrichment for three of the four TFs with the exception of EcR. These results imply that genes in cluster 3 are those that need specific regulation of their expression during development, as the studied TFs are implicated in developmental pathways. In contrast to cluster 3, cluster 1 is the one most depleted in TF binding, as expected since it is the cluster with the smallest proportion of active genes.

When observing the average Pol II profile for each cluster one of the main distinctions is the location of the Pol II peak, which is shifted dramatically upstream in cluster 4. To see whether the TF binding data reflects this shift, I plotted the median distance between the TSS and the different TFs for each cluster in Figure 4.30b. For all four TFs, cluster 4 shows a shift upstream compared to the other clusters, mirroring the shift of Pol II. Together, these results show that not only does the enrichment of TF binding change between clusters, but also the difference in the location of the Pol II peak of cluster 4 compared to the TSS is accompanied by a change in the location of the TF binding sites.
4.2.7 Nucleosome positioning

The positioning of nucleosomes around promoters plays an important role in transcription regulation. To explore the relationship of nucleosomes with the different groups of promoters, I downloaded MNase-seq data from Gilchrist et al. (2010), in which chromatin is digested by micrococcal nuclease, leaving the fragments of DNA which are protected by a nucleosome. After mapping the paired-end reads to the *D. melanogaster* genome, I used the GeneTrack software package (Albert et al., 2008) to identify the nucleosome locations. GeneTrack smooths the mapped data using a Gaussian distribution and identifies non-overlapping peaks that do not fall within a user-defined interval of each other. Given that nucleosomes have a natural 147 bp exclusion zone determined by the wrapping of the DNA around them, the interval parameter of GeneTrack was set to this value. In GeneTrack the height of the peaks corresponds to the estimated number of reads present at their location.

Nucleosomes show a distinct occupancy pattern around transcription start sites, with a very precise positioning for the first nucleosome downstream of the TSS (the +1 nucleosome), followed by other nucleosomes at positions that get less well defined the further away they are from the TSS. When the average nucleosome occupancies around transcription start sites are plotted in Figure 4.31a a clear periodic pattern of nucleosome occupancy appears after downstream of the TSS.

The unoccupied area between the nucleosome just downstream of the TSS (+1 nucleosome) and the first nucleosome upstream (-1 nucleosome) is called the Nucleosome Free Region (NFR). This is the area accessible to DNA-binding proteins without any nucleosome remodelling and it can vary between promoters. I ranked the promoters by their expression status and plotted their NFR size in Figure 4.31b. There is a marked, if small, decrease in NFR size between expressed and non-expressed promoters, from around 180 bp to 150 bp, but no change in NFR with expression level inside the group of expressed promoters. The expression values taken from the Affymetrix GeneChip of all genes have a bimodal distribution (not shown), and the valley between the two modes was taken as the threshold above which a gene is classified as expressed for this analysis.

Apart from the nucleosome positioning and spacing, we can define other measures for the nucleosomes. One of them is their “focus”, meaning how precise their location is across the population of cells. The height of the peak detected by GeneTrack reflects both the number of reads used to define the position of a nucleosome and how dispersed those reads are: if they all map to the same location, the height of the peak will approximate the number of reads, while if the reads are close, but not in the same location, the peak will be lower. As a measure of nucleosome focus, I used the height of the GeneTrack peak divided
RNA POL II HAS DISTINCT MODES OF OCCUPANCY IN FLY

Figure 4.31: (a) Average nucleosome occupancy across all genes. There is a distinct periodic pattern downstream of the TSS and into the gene body, with a well positioned +1 nucleosome. The peak corresponding to the -1 nucleosome is much less pronounced. (b) There is a slight decrease in the NFR size between expressed and non-expressed genes. The lines represent the sliding quartiles of NFR size. The genes were ranked by decreasing expression value. The threshold between expressed and non-expressed genes is marked by the red vertical line.

by the number of reads that overlap the determined nucleosome location, as was done in previous nucleosome studies (Zaugg and Luscombe, 2011).

The focus of both the +1 and -1 nucleosomes shows a very marked difference between expressed and non-expressed promoters. After ranking the genes by expression values, I calculated the quartiles over a sliding window of 100 promoters. The different quartiles of this sliding window for both the +1 and -1 nucleosomes are plotted in Figure 4.32. The focus of both nucleosomes is relatively constant across the expressed genes, but there is a strong decline in focus between expressed and non-expressed ones. These results show that the positioning of nucleosomes in expressed genes is more precise.

To see if there is any association between Pol II profiles and nucleosome positioning, I plotted the average nucleosome occupancy for the transcripts of each cluster. If the shift in the Pol II peak from clusters 4 and 5 is reflected on the nucleosome level, we would expect to see a shift also in the position of the +1 and -1 nucleosomes. As seen in Figure 4.33a, the average position of both nucleosomes is similar for all clusters.

Given that the clusters include a mix of expressed and not expressed transcripts, I also plotted the occupancies separately for the transcripts that are expressed or not when assayed by microarray profiling (Figure 4.33b, left panel) and for the more stringent expressed/stalled cate-
4.2 results

Figure 4.32: Nucleosomes from expressed genes are well positioned compared to those from non-expressed genes. The lines represent the quartiles on a sliding window of 100 promoters of the -1 (a) and +1 (b) nucleosome focus. The genes were ranked by decreasing expression value. The IQR is shaded in grey. The threshold between expressed and non-expressed genes is marked by the red vertical line.

categories based on both microarray and Pol II data defined in section 4.2.3 (Figure 4.33b, right panel). Only in stalled genes is the +1 nucleosome displaced. For all other situations, the nucleosomes positions are similar regardless of the observed shift in the Pol II peak.

All in all these results indicate that the differences in Pol II and TF positioning are not reflected on nucleosome positioning.

4.2.8 Histone modifications and variants

As seen in the previous section, the positioning of the +1 and -1 nucleosomes is not changed between the different clusters. However, there are other variables at play when it comes to nucleosomes, including the posttranslational modification of histones and whole histone variants. To see whether the clusters have different histone modifications or variants I downloaded datasets of chromatin modifications and variants from modENCODE [Kharchenko et al., 2010]. The datasets were for the following variants and modifications: H2A.V, H2BK5 acetylation, H2B ubiquitination, H3K18 acetylation (H3K18ac), H3K23ac, H3K27ac, H3K27 dimethylation (H3K27me2), H3K27 trimethylation (H3K27me3), H3K36 monomethylation (H3K36me), H3K36me3, H3K4me, H3K4me2, H3K79me, H3K79me2, H3K9ac, H3K9me2, H3K9me3, histone H4, tetraacetylated histone H4 (H4acTetra), H4K16ac, H4K20me, H4K5ac and H4K8ac. A dataset for Heterochromatin Protein 1 (HP1a) was also included. Among the datasets used
Figure 4.33: (a) The position of the +1 nucleosome is identical in all clusters, despite the different location of Pol II. (b) Active genes (red line) have more defined nucleosome positioning than non-transcribed ones (green line). In addition, the +1 nucleosome is shifted upstream into the TSS for stalled genes (purple line).
are modifications associated with transcription activation (H3K27ac and H3K4me2, for example) and with repression (H3K9me3 and H3K27me3, for example).

I used the processed data provided by modENCODE which details the genomic regions where there is significant signal from the ChIP experiment for each histone variant or modification. I identified which of these regions are within 400 bp of the TSSs. I then calculated the enrichments of the histone variants and modifications among the Pol II clusters in the same way as for the TF analysis in section 4.2.6, using equation [1]. Given that the data from modENCODE are from the S2 cell line, I used the cluster classification for that cell type.

Some patterns emerge when considering the enrichments or depletions for each mark that the clusters as a whole have (Figure 4.34a). For example, marks found at enhancers and near promoters like H3K27ac or H3K4me2 are seemingly depleted in cluster 1, and there is a strong depletion of repression-associated H3K27me3 in clusters 2 to 5. However, these patterns are merely the result of the higher proportion of expressed genes in clusters 2 to 5 compared to cluster 1. When the enrichments for the subsets of each cluster classified as expressed or not expressed based on microarray data are calculated (figure 4.34b), the differences between the clusters are reduced. Thus the main signal comes from the expression status of the gene and not the Pol II profile. For example, H3K27me3 is strongly depleted in expressed genes in all clusters, and slightly enriched in unexpressed genes.

Thus, as with nucleosome positioning, individual chromatin marks at promoters do not appear to differ among the Pol II clusters.

4.2.9 Chromatin types

In the previous section, individual chromatin marks do not show a preference for the different clusters as transcription factors do. With the expansion of available data new computational techniques have been used to capture the patterns of binding across different chromatin datasets. In one such example, Filion et al. (2010) combined extensive DamID datasets to separate the *D. melanogaster* genome in five chromatin states, or “colours”, which reflect the protein makeup of each area.

DamID is a technique that, like ChIP-seq, identifies genomic regions that interact with specific proteins (van Steensel et al., 2001). ChIP-seq relies on the enrichment of DNA sequences by cross-linking them to proteins and immuno-precipitating the protein of interest using antibodies. DamID, on the other hand, works by expressing a construct of the protein of interest fused with the *Escherichia coli* DNA adenine methyltransferase (Dam). This construct methylates the DNA around the binding sites of the protein of interest. Afterwards, DNA is treated with the DpnI restriction enzyme that cuts methylated GATC sequences, creating small DNA fragments that correspond to the bound
84 RNA POL II HAS DISTINCT MODES OF OCCUPANCY IN FLY

Figure 4.34: (a) Enrichments of different histone modifications and variants for the different clusters. (b) Enrichments of the same histone modifications and variants for the subsets of each cluster defined as expressed or not expressed by microarray data.
regions. These are then sequenced or hybridised to a tiling array. While the resolution of this technique is constrained by the presence of GATC sequences in the genome, it has the advantage over ChIP-seq that it does not need an antibody specific for the protein of interest. Another limitation of the DamID technique is that it requires the expression of a fused protein that is not endogenous to the organism. The increased levels of this protein can lead to binding patterns or other biological behaviours that would not be observed in the wild type.

In this study, Filion and colleagues created binding maps for 53 different nuclear proteins and combined them using a hidden Markov model into five different states. The five colours are:

- green, a repressed state marked by a presence of HP1;
- black, a repressed state;
- blue, a state repressed by Polycomb;
- yellow, marking transcribed house-keeping genes
- red, marking transcribed genes with restricted expression patterns and linked to specific cellular processes.

If the differences in TF binding at the different clusters reflects differences in regulation, and in particular a presence of development-linked genes in cluster 3 compared with the other cluster of expressed genes, cluster 2, these differences should be reflected in the chromatin colours present at each TSS. To find out if this is the case, I counted the number and colour of segments that overlap with each TSS. Those counts are represented in Figure 4.35 for the subsets of expressed and stalled transcripts in each cluster. Among the expressed transcripts, cluster 3 has a higher proportion of TSSs that fall within a red region (i.e.: marks for genes with specific, more regulated functions). Cluster 2, also with a high number of expressed genes, has less developmental TF binding sites and is marked with the yellow colour characteristic of more widely expressed genes. Among the stalled transcripts, most show the black or blue repressive colours, as would be expected. The relative higher proportion of transcripts with TSSs overlapping red regions in cluster 3 is in agreement with the enrichment of developmental TF binding sites, suggesting that this cluster contains genes which need to be expressed at the right times in development.

4.2.10 Go analysis

As seen previously, the data from TF binding and chromatin colours suggest that of the two clusters that have a highest proportion of expressed transcripts, cluster 2 represents less regulated, more widely expressed transcripts, while cluster 3 represents transcripts with higher TF binding and more focused expression. To see if these results are
supported by Gene Ontology annotations I ran G:Cocoa, part of the G:Profiler web interface (Reimand et al., 2007, 2011). G:Cocoa takes as its input several lists of gene IDs and compares their GO annotations.

I ran G:Cocoa separately for expressed and stalled genes. The results between the two are mainly in agreement with each other and with the suggestions from the TF binding and chromatin colour analysis. Cluster 3 shows an enrichment for terms related to developmental processes, such as organ development (GO:0048513), appendage morphogenesis (GO:0035107) and nervous system development (GO:0007399). Cluster 2 includes some terms related to developmental processes though to a lesser extent, and in addition is enriched in terms related to cytoskeleton organisation (GO:0007010) and cell death (GO:0008219), among others. Among stalled genes, cluster 5 shows enrichment for terms related to nucleic acid metabolic processes (GO:0090394) and cell cycle processes (GO:0022402).

The results for clusters 2 and 3 support the hypothesis that the first represents genes involved in general processes while the latter represents genes involved in more regulated and cell-type specific processes. With a high proportion of stalled genes, cluster 5’s enrichment for cell cycle processes may be reflecting the fact that salivary gland cells are no longer dividing.

4.2.11 Comparison with previous classifications

Previous studies have classified D. melanogaster promoters based on their underlying DNA sequence and studied how those categories relate to the gene’s functions, chromatin modifications and even the behaviour of Pol II (Engström et al., 2007; Ni et al., 2010; Rach et al., 2011). In
this section I compare the characteristics of all categories to try to find correspondences between them.

Engström et al. (2007) used the dataset of core promoter models presented by Ohler (2006) and used by me in section 4.2.5. In that work, the authors find that there is an enrichment of developmental genes among those whose core promoter types contain the Inr element (Inr only, Inr/DPE or TATA/Inr). Cluster 3 also shows the highest proportion of core promoter types with the Inr element (see table 11) and an enrichment of developmental genes (see section 4.2.10).

Rach et al. (2011) used a classification defined by Ni et al. (2010). In this work the distribution of sequenced capped RNA fragments was used to classify promoters. Three classes were identified: broad distribution of reads with a distinct peak in the middle (BroadPeak), narrow distribution of reads around a peak (NarrowPeak) and a broad, weak peak (WeakPeak). The overlap between these three categories and the five clusters I’ve identified are present in table 12. For all of the clusters the highest category of three is the WeakPeak, with clusters 2, 4 and 5 having the largest proportion of WeakPeak patterns among their promoters. Clusters 1 and 3 are the clusters with the highest proportion of NarrowPeak patterns. All clusters have similar levels of BroadPeak patterns.

In their analysis of both D. melanogaster 0-24 hour embryos and H. sapiens CD4+ cells, Rach et al. (2011) find that NarrowPeak promoters have the highest proportion of TATA box core promoter motifs and also show a higher presence of Pol II. In agreement with the overlaps found, NarrowPeak promoters would then be analogous to cluster 1 and 3. These two clusters show the highest proportion of overlap with the TATA-box core promoter element and cluster 3 has the most pronounced Pol II peak in the datasets used here.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>BroadPeak</th>
<th>NarrowPeak</th>
<th>WeakPeak</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13.74%</td>
<td>39.74%</td>
<td>46.52%</td>
</tr>
<tr>
<td>2</td>
<td>15.08%</td>
<td>16.30%</td>
<td>68.62%</td>
</tr>
<tr>
<td>3</td>
<td>17.07%</td>
<td>28.91%</td>
<td>54.01%</td>
</tr>
<tr>
<td>4</td>
<td>16.39%</td>
<td>13.44%</td>
<td>70.16%</td>
</tr>
<tr>
<td>5</td>
<td>18.92%</td>
<td>11.45%</td>
<td>69.63%</td>
</tr>
</tbody>
</table>

Table 12: Correspondence between Pol II clusters and gene patterns presented in (Ni et al., 2010).

Interestingly, BroakPeak and WeakPeak promoters, show a more ordered nucleosome organisation than NarrowPeak promoters. In the analysis of my clusters no strong differences in nucleosome organisation were observed between the clusters. A lack of clear correspondence in this case might be due to the difference between the cell types used in the two studies.
4.3 Discussion

In this chapter I identified five different profiles of RNA Pol II binding at TSSs via an unbiased clustering approach. While the biological mechanism that leads to these profiles is not clear, they show distinct expression behaviour, underlying regulatory sequences, transcription factor binding patterns and functional annotations. On the other hand, these behaviours are not reflected in the nucleosome architecture or composition at the studied TSSs.

Clusters 4 and 5 are distinct in their relative abundance of stalled genes and, more markedly, in the upstream shift in the Pol II profile at their TSSs. Remarkably, this shift is not reflected in the positioning of the respective +1 nucleosomes, but it is reflected in the positioning of TFs binding to these promoters. The stalled genes of cluster 5 from salivary glands are enriched in cell cycle processes which are not active in this tissue. These two clusters are also those that show more change between salivary gland and S2 cell samples. It is possible that they are reflections of repressive regulation.

Among the five clusters, clusters 2 and 3 stand out in their high proportion of expressed genes. Cluster 2 has a broad peak of Pol II around the TSS, similar to the average of all the transcripts. The chromatin in those TSSs is enriched in a state associated with broad expression pattern across many tissues and the transcripts have Gene Ontology annotations related to ubiquitous processes. Cluster 3 has a sharper profile, more centred on the TSS. In terms of underlying sequence, it has more promoters of the Inr and Inr/DPE types than any other cluster. In addition, TSSs belonging to this cluster have higher variability in expression during fly development, are enriched for TF-binding, are more often located in chromatin areas associated with specific cellular processes and are enriched in GO terms for morphogenesis processes. I hypothesise that the different Pol II shape is reflecting a different mode of regulation between broadly-expressed and tightly regulated genes.

Even after establishing that some of the clusters are associated with developmentally regulated genes, I still lack a mechanistic explanation of the different observed shapes. There are several possible options. Two main differences in the Pol II signal shape are observed: the sharpening of the peak of cluster 3 compared to cluster 2, and the upstream shift of the peak of clusters 4 and 5 compared to the others. For both cases there is no difference in nucleosome positioning, so I cannot assume that it is they who restrict the position of Pol II.

For the sharpening of cluster 3’s profile compared to cluster 2, possible explanations might lie in the higher TF binding and in the link to development. If these transcripts need a more tightly regulated expression, their recruitment might in turn be more tightly regulated, and the transcription factors that participate might constrain the location of Pol II. Cluster 2, with its broader profile, could be seen as anal-
ogous to vertebrate CpG promoters, which have broader TSSs. Promoters of this type have been associated with the DPE element in *D. melanogaster*, where a broader Pol II profile is also reported (Rach *et al.*, 2011).

The upstream shift of the Pol II peak in clusters 4 and 5 has no evident explanation from the analysis carried out. The fact that this shift is also observed in the TF binding location indicates that the shift is present for other chromatin components. While some of the shift is indubitably accounted for by the higher presence of stalled genes in these clusters, it should be stressed that this shift is still observed for the expressed genes present in the clusters. Perhaps these TSSs are alternative forms of other transcripts which start just upstream, or they might be an hitherto uncharacterised type of promoter. As this analysis is sensitive to the TSS positions, in further work these could be defined using experimental datasets such as CAGE, so that the foundation for the work would be based on TSSs that had been experimentally validated as being active in the conditions being studied and their coordinates more precisely mapped.

The addition of data from chromatin remodellers and basal transcription factors could help elucidate the reason why Pol II is seen in different locations on these promoters. Special attention should be paid so that the data used is in accordance with the cell types and developmental stages that were used to classify promoters according to their Pol II profiles since, as was seen when comparing the salivary gland and S2 samples, promoters can change Pol II classes.
The chromosome conformation capture data used in this chapter was created as part of a collaboration between the Fraser Laboratory of the Babraham Institute, the Osborne Laboratory of the same Institute and the Luscombe Laboratory. The promoter capture-Hi-C technique was developed by Cameron Osborne from the Osborne Laboratory and Stefan Schoenfelder from the Fraser Laboratory. The human promoter capture-Hi-C libraries were created by Cameron Osborne and the mouse libraries were created by Stefan Schoenfelder and Mayra Furlan-Magaril, also from the Fraser Laboratory. The raw data was processed by Bori Mifsud and Robert Sugar from the Luscombe laboratory to identify the significantly interacting fragments that I used in my analysis. Bori Mifsud also provided the expression classes for human and mouse genes.

5.1 Introduction

5.1.1 3C-derived techniques capture chromosome conformation

Since the publication of the original Chromosome Conformation Capture (3C) technique by Dekker et al. (2002) there has been progress in the number of derivative protocols to capture the three-dimensional structure of DNA in living cells. The basis of all techniques relies on cross-linking the nuclear content and cutting the DNA with a restriction enzyme. This is followed by ligation of the ends of fragments in diluted conditions to increase the chances of interaction between those fragments that were spatially close in the nucleus. After reversing the cross-linking, the experimenter is left with DNA molecules comprising two different genomic regions.

The different 3C-derived techniques diverge from each other in the way the ligated fragments are detected. The original 3C uses PCR amplification to test whether two queried sites interact in the nuclear space. The first technique to build upon 3C was Chromosome Conformation Capture-on-Chip (4C; Simonis et al., 2006), which uses specific PCR primers to amplify the ligation fragments that interact with a chosen locus. The amplified fragments are then hybridised to a chip (in the original technique) or sequenced. As a result, this technique identifies all the regions in the genome that interact with the probed locus. Compared with 3C, which tests the interaction between two individual loci — it’s a "one-to-one" technique —, 4C is a "one-to-all" technique. The
logically named 3C-Carbon Copy (5C; Dostie et al., 2006) uses pairs of oligonucleotides that contain sequences complementary to the fragments that are to be queried. These pairs of oligonucleotides are used to identify the pairwise interactions between the queried fragments. This way, it complements 3C and 4C as a "many-to-many" technique.

A more unbiased genome-wide technique called Hi-C was developed in 2009 (Lieberman-Aiden et al., 2009). Hi-C marks ligation sites with biotin and uses a streptadivin pull-down to enrich the library in fragments that are the result of successful ligations, improving the efficiency of the original 3C technique. The biotin-enriched library is then sequenced using paired-end sequencing. After mapping both ends to the reference genome it is possible to identify pairs of interacting regions in the entire genome. While a great advance for unbiased conformation capture, the vast number of possible interactions inside an entire genome means that for equal sequencing depth Hi-C has lower resolution than techniques like 4C or 5C. Lieberman-Aiden et al. (2009), for example, created an interaction map with a spatial resolution of 1 Mb. This means that we can only assert that there is an interaction between two large regions of the genome. We do not know precisely what elements within those regions are interacting.

A final variation of 3C is Chromatin Interaction Analysis by Paired-End Tag sequencing (ChIA-PET; Fullwood et al., 2009). It is a directed version of 3C that uses a protein-DNA interaction as its selection criterion. It reduces the complexity of the ligation library using an antibody against a protein of interest. This way, the fragments assayed by sequencing are enriched for interactions in which at least one of the ends is cross linked with the protein.

5.1.2 Different techniques for different questions

The different pros and cons of the 3C-derived techniques make each of them appropriate to answer different questions. Hi-C is suited to identify the overarching genome organisation. The main metazoan model organisms (yeast, fly and mouse) and human have been assayed using Hi-C (Duan et al., 2010; Sexton et al., 2012; Dixon et al., 2012), where contiguous regions in the chromosomes which were called Topologically Associated Domains (TADs) were identified. The regions inside TADs are very interconnected, while interactions between domains are rarer.

More directed techniques are used when a specific protein or process is being investigated and the increase in resolution is advantageous. The original ChIA-PET study identified the interaction network bound by oestrogen-receptor-α and found evidence that it brings together promoters for co-regulation (Fullwood et al., 2009). More recently, Handoko et al. (2011) used ChIA-PET to characterise the interaction network of CTCF, an insulator protein involved in chromatin looping.
Techniques directed at specific loci like 5C was used to probe the \( \beta \)-globin Locus Control Region (Dostie et al., 2006) or the promoters and enhancers of the 1% of the human genome studied in the pilot ENCODE project (Sanyal et al., 2012).

The specific interactions involving oestrogen-receptor-\( \alpha \) or CTCF or those in the \( \beta \)-globin locus might be lost in the all-encompassing Hi-C due to its lower resolution. On the other hand, ChIA-PET, 5C and 4C lack the genome-wide property of Hi-C that allows the discovery of new interactions in other genomic regions. Until the sequencing capabilities improve and lead to an increase in the resolution of 3C-derived techniques, there will be a trade-off between the amount of the genome being interrogated and the resolution at which individual interactions can be assayed.

5.1.3 The promoter capture-Hi-C technique

With this trade-off, Cameron Osborne and Stefan Schoenfelder developed an extension of Hi-C to enrich for interactions involving promoters, called Capture-Hi-C (C-Hi-C). Compared with 4C, which queries all interacting regions of a single locus, or Hi-C, which attempts to query all the interacting regions of the entire genome, promoter C-Hi-C identifies all interacting regions of promoter regions. Thus it lies between the two in the trade-off of 3C-derived techniques. The concept behind this new technique is similar to that of other sequence capture techniques. First, we create a set of biotinylated RNA baits that targets promoters in the organism of interest. The baits are then used to enrich a ligation library created in the same manner as for Hi-C. This way, we obtain a ligation library enriched in ligation fragments from interactions involving promoters.

Here, using the the list of interactions created from the first genome-wide promoter C-Hi-C experiments, I investigated the chromatin landscape of promoter-interacting regions in murine ES cells, murine fetal liver, the human laboratory cell line GM12878 and human CD34+ cells.

5.2 Results

5.2.1 General overview of the promoter C-Hi-C data

We generated libraries of C-Hi-C for four different cell types over two mammalian species. Two biological replicates were made from \textit{Mus musculus} embryonic stem cells (mESC) and from fetal liver samples taken at embryonic day 14.5. For \textit{Homo sapiens} three replicates were created from the GM12878 cell line and two from CD34+ cells.

The genomes were divided into the fragments created using the Hind III restriction enzyme. This resulted in 823,377 fragments in \textit{M. musculus} and 837,161 in \textit{H. sapiens}. Of these, 22,266 and 21,841, respectively,
hybridise to the set of promoter baits used during the enrichment step of the C-Hi-C protocol. In the following analysis I refer to these fragments as bait fragments, to interactions involving two bait fragments as promoter-promoter interactions and to interactions involving a bait fragment and a non-bait fragment as promoter-other interactions.

The method used to call significant interactions assumes that the relative sequencing coverage of a genomic region reflects the biases that the region might have (GC-content, mappability). With this assumption, the probability of a random interaction between any two fragments of the genome is the product of their relative coverages and the noise component (set to a conservative value of 1 in our analysis). Our datasets contain not only the relative coverages of all fragments, but also the actual number of observed interactions between any two fragments in the number of paired-end reads that map to distinct fragments. The probability of observing this number of reads in our dataset by chance is given by a cumulative binomial. This results in a p-value for the likelihood of observing at least that number reads for a given interaction. After correcting for multiple testing using the Benjamini-Hochberg multiple testing correction, we obtain a q-value which is used to infer significant interactions for a given false discovery rate.

Interactions were called independently for each sample, for a false discovery rate of 0.05. Significant interactions that appeared in two or more replicates were taken as the final set of interactions. Finally, interactions between a bait fragment and non-bait fragment that were separated by less than 20 kb were not considered in the analysis.

The numbers of promoter-other interactions are present in table 13, as well as the number of baits and individual non-bait fragments involved. All conditions contain interactions involving most of the bait fragments. At more than 700,000, fetal liver has the largest number of interactions of the two mouse samples, compared with around 550,000 in the mESCs. In the human samples, the GM12878 samples have over 720,000 interactions and CD34+ cells have 360,000 interactions.

<table>
<thead>
<tr>
<th>Species</th>
<th>Sample</th>
<th># int. baits</th>
<th># int. frags.</th>
<th># interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>M. musculus</td>
<td>mESC</td>
<td>18,530</td>
<td>284,233</td>
<td>545,054</td>
</tr>
<tr>
<td>M. musculus</td>
<td>Fetal liver</td>
<td>19,653</td>
<td>299,782</td>
<td>708,899</td>
</tr>
<tr>
<td>H. sapiens</td>
<td>GM12878</td>
<td>19,825</td>
<td>348,893</td>
<td>721,472</td>
</tr>
<tr>
<td>H. sapiens</td>
<td>CD34+</td>
<td>19,014</td>
<td>224,926</td>
<td>360,118</td>
</tr>
</tbody>
</table>

Table 13: Number of promoter-other interactions and baits involved

As seen in the histograms of figure 5.36a, the number of individual fragments that a promoter interacts with can vary widely, with some exceptional promoters interacting with more than 100 fragments. However, the fragments that contact these highly-interacting promoters are not necessarily highly-interacting themselves. This can be seen in figure 5.36b for the mESC sample, where I plot every interaction according
to the number of fragments the promoter of that interaction contacts and the number of promoters the fragment of that interaction contacts. If highly-interacting promoters contacted highly-interacting fragments, there would be a positive correlation between the two. Instead, there is an inverse relationship, implying that hub promoters contact less connected fragments and vice-versa. An identical result is obtained for the other three samples.

![Graphs showing number of interactions and number of bait fragments](a)

![Smoothed scatter-plot](b)

Figure 5.36: (a) Number of fragments a bait fragment interacts with, for each of the available samples. (b) Smoothed scatter-plot of the number of fragments a bait interacts with on the X-axis and the number of baits a fragment interacts with on the Y-axis, for the mESC sample.

Promoter-promoter interactions were called using a variation of the binomial technique that uses logit regression to determine interactions and estimates the number of false-positives from random ligation samples. Promoter-promoter interactions called by this technique on the same samples as the promoter-other interactions are listed in table 14. Due to lower coverage from the human random ligation samples, the number of promoter-promoter interactions called is much lower than in *M. musculus*, with over one-million promoter-promoter interactions called in mESCs compares to less than one quarter of a million in GM12878 cells.

<table>
<thead>
<tr>
<th>Species</th>
<th>Sample</th>
<th># int. baits</th>
<th># interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>M. musculus</em></td>
<td>mESC</td>
<td>20,349</td>
<td>1,139,849</td>
</tr>
<tr>
<td><em>M. musculus</em></td>
<td>Fetal liver</td>
<td>20,382</td>
<td>797,611</td>
</tr>
<tr>
<td><em>H. sapiens</em></td>
<td>GM12878</td>
<td>19,899</td>
<td>235,534</td>
</tr>
<tr>
<td><em>H. sapiens</em></td>
<td>CD34+</td>
<td>19,718</td>
<td>118,571</td>
</tr>
</tbody>
</table>

Table 14: Number of promoter-promoter interactions and baits involved
5.2.2 Expression levels in the studied conditions

We assigned expression classes to the bait fragments for all four samples using corresponding publicly available RNA-seq datasets. Each promoter was put into a different category according to expression status of its gene. For mouse, mESC data was taken from Shen et al. (2012) and fetal liver from Kowalczyk et al. (2012). The RNA-seq datasets for both human samples were taken from the ENCODE project (Djebali et al., 2012).

If the RPKM value (reads per kilo base per million mapped reads) of the gene is zero, its promoter is assigned to the 0 expression class. The other promoters are sorted according to the RPKM values of their genes and assigned to the four quartiles (with 1 being the promoters with lowest expression and 4 those with the highest). Bait fragments can overlap with more than one promoter. If all the promoters that map to the same bait fragment have the same level of expression, that is the level assigned to the fragment. If they are not, the bait expression category is marked as not available (NA). The number of bait fragments that fall in each expression class is given in Table 15.

<table>
<thead>
<tr>
<th>Expression class</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>NA</th>
</tr>
</thead>
<tbody>
<tr>
<td>mESC</td>
<td>3533</td>
<td>3753</td>
<td>3526</td>
<td>3422</td>
<td>3252</td>
<td>4739</td>
</tr>
<tr>
<td>Fetal liver</td>
<td>7573</td>
<td>2761</td>
<td>2486</td>
<td>2410</td>
<td>2375</td>
<td>4620</td>
</tr>
<tr>
<td>GM12878</td>
<td>6582</td>
<td>2168</td>
<td>1769</td>
<td>2697</td>
<td>3216</td>
<td>5409</td>
</tr>
<tr>
<td>CD34+</td>
<td>2945</td>
<td>1871</td>
<td>1888</td>
<td>2989</td>
<td>3328</td>
<td>8820</td>
</tr>
</tbody>
</table>

Table 15: Number of bait fragments in each expression class.

5.2.3 Highly-expressed promoters interact with annotated enhancers

The design approach for our baits leads to a bias of interactions involving promoters, but not involving other regulatory sequences. Enhancers are classic long-distance regulatory sequences that are implicated in activation of transcription by looping and contacting the promoter region. In order to evaluate the extent to which promoters interact with enhancers we integrated our interaction data with data for gene expression and enhancer location.

We incorporated enhancers in our analysis using several available datasets for our different cell types. All the set of enhancers were identified via the presence of H3K4me1, which is known to localise to these distal regulatory regions. Enhancers from mESC and mouse fetal liver were taken from Shen et al. (2012), which identify regions that have high levels of H3K3me1 and a low level of H3K4me3. For the human cell line GM12878, enhancers were taken from the genome segmentation analysis of ENCODE (ENCODE Project Consortium et al., 2012), which classified areas of the genome according to their protein.
Figure 5.37: Proportion bait fragments in each expression class that interact with different number of enhancers, for the mouse (a) and human (b) samples. The proportion of non-expressed genes (in yellow) increases with the number of enhancers that interact with the bait fragments.

5.2.4 Histone modifications in interacting regions reflect expression status

The nucleosomes of promoters and other regulatory sequences are known to have certain post-translation modifications. To determine enhancers, for example, Shen et al. (2012) selected regions with H3K4me1 but not H3K4me3, which in turn is enriched at promoters. Our survey of promoter-contacting regions is not biased towards enhancers, so I char-
acterised the pattern of chromatin modifications in all fragments that interact with promoters.

To investigate the effect that the transcriptional status of a promoter might have on the chromatin composition of its interacting fragments, I separated fragments according to the expression level of the promoters they interact with. Given that a fragment can interact with promoters in different expression classes, I reduced the set of fragments to those that interact only with promoters of the same expression class. The number of such fragments for each sample and expression class can be seen in table 16. While this reduced the set of fragments available for the analysis, it assures that there is no ambiguity in the classification of interacting fragments.

<table>
<thead>
<tr>
<th>Expression class</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>mESC</td>
<td>27,524</td>
<td>38,842</td>
<td>30,398</td>
<td>24,654</td>
<td>16,062</td>
<td>137,480</td>
</tr>
<tr>
<td>Fetal liver</td>
<td>70,763</td>
<td>20,528</td>
<td>16,664</td>
<td>14,054</td>
<td>11,149</td>
<td>133,158</td>
</tr>
<tr>
<td>GM12878</td>
<td>79,586</td>
<td>22,333</td>
<td>16,850</td>
<td>20,580</td>
<td>25,445</td>
<td>164,794</td>
</tr>
<tr>
<td>CD34+</td>
<td>22,889</td>
<td>15,708</td>
<td>13,526</td>
<td>21,932</td>
<td>19,969</td>
<td>94,024</td>
</tr>
</tbody>
</table>

Table 16: Number of fragments that interact with only one expression class.

A variety of histone modifications were used in this analysis from different sources. As the GM12878 is a tier-1 ENCODE cell line, a wide variety of datasets are available for it. Those datasets were used for this cell line, using the ENCODE peak-calling pipeline. For all other samples where the raw data was available, I mapped the reads to the reference genome (mm9 in the case of *M. musculus* samples and hg19 in the case of *H. sapiens* samples) using Bowtie (Langmead et al., 2009), allowing for one mismatch in a seed of 25 bp and filtering reads that map to more than one location. Peaks were called using the MACS software at default settings (Zhang et al., 2008). When raw data was not available, I took the peaks called in the original publication. For a list of datasets used see tables 19, 20, 18 and 17.

Using the peaks called for the different histone modifications, I classified fragments as having a modification if they overlapped with the peak summit location. Then, for each histone mark, I calculated the enrichment using equation (2), which determines how more likely a fragment that interacts with a promoter of a certain expression class is to have the mark compared with all non-bait fragments. The result has a positive value if the the fragments of an expression class are enriched in a mark compared to all non-bait fragments and a negative value if they are depleted. Since the promoter regions in bait fragments are known to be enriched for histone modifications, non-bait fragments are used instead of all fragments as the reference so as not to skew the analysis.

\[
enrichment = \log_2 \left( \frac{\# \text{Fragments In Expression Class Overlapping Mark}}{\# \text{Fragments In Expression Class}} \right) \cdot \frac{\# \text{Non–bait Fragments Overlapping Mark}}{\# \text{Non–bait Fragments}}
\]
Figure 5.38: Enrichments of interacting fragments overlapping with different modifications over all non-bait fragments, for (a) mESC, (b) fetal liver, (c) GM12878 and (d) CD34+ cells. Fragments interacting with promoters in the higher expression categories are enriched in marks for active transcription compared with fragments that interact with lowly- or non-expressed promoters.
Having calculated the enrichment or depletion of histone marks for all four samples I plotted them in figure 5.38 (yellow and red denote positive enrichment values while blue denotes negative values). A striking pattern that is immediately clear is that fragments interacting with more highly expressed promoters (rows marked 3 and 4) are enriched for most marks, while fragments interacting with non- or lowly-expressed promoters (rows 0 through 2) are depleted for the same marks or show smaller enrichments. This result highlights how areas of the genome that interact with active promoters have more active chromatin themselves. Further, the clear exception is the repressive mark H2K27me3 (marked in red), for which the pattern is inverted, with fragments that interact with highly expressed promoters being depleted compared with fragments that interact with less active promoters.

Figure 5.39: Enrichments for chromatin marks of interacting fragments that contain (or not) enhancers, for the mESC (a) and fetal liver (b) mouse samples. Even in non-enhancer fragments there is an enrichment in activating marks for the fragments that interact with highly expressed promoters.

Enhancers are characterised by the presence of histone marks such as H3K4me1 and H3K27ac. As such, the enrichment for these marks that is seen in our datasets for fragments interacting with active promoters could be due to an enrichment of enhancers in those fragments. As shown in figure 5.37, active promoters tend to interact with more enhancers. To see if this is the case, I separated the fragments that contain enhancers from those that don’t, and calculated the enrichments of the chromatin marks for the two sets. Those results are shown for the mouse samples in figure 5.39. There is a clear difference in the overall enrichment between enhancer and non-enhancer fragments, as expected given that histone modifications were used to identify these regions. This is particularly true for H3K4me1 and H3K27ac, the marks used to define the set of enhancers used here and hence heavily enriched.
in enhancer fragments. However, even for non-enhancer fragments the observation that fragments that interact with highly expressed promoters have more activating chromatin marks is still true.

5.2.5 **Histone modification enrichment pattern is robust with regards to interaction distances**

A caveat of the previous analysis of histone modifications in fragments interacting with promoters is that the observed enrichments and depletions be a result of very short range interactions. If that is the case, the enrichments are merely a reflection of the histone modifications present at the gene promoter or body.

In order to control for this potential bias in the results, I performed the same analysis on the subset of interactions that span over a set distance threshold. The thresholds used were 100,000, 500,000 and 1,000,000 base pairs. The enrichments for these interaction subsets in each cell type and organism are plotted in figures 5.40 and 5.41 for the mouse and human samples, respectively. For all four samples, as distance increases, the enrichments or depletions for most histone modifications become more pronounced. This maintains the same pattern of enrichment seen in the full dataset, regardless of the distance threshold used: histone marks associated with transcription are enriched in fragments that interact with highly expressed promoters compare with fragments that interact with lowly or non-expressed promoters, while histone marks associated with repression show the inverse pattern.

A similar control for the set of interacting fragments that overlap or not with annotated enhancers is shown in figure 5.42, using the same distance thresholds as in the previous paragraph. As with the previous control, the patterns of histone modification enrichments are maintained in the interaction subsets that fall above the distance thresholds used.

5.2.6 **ChromHMM analysis**

To explore further the relationship between interacting fragments and histone modifications I added information regarding the chromatin state. Chromatin states have been defined using the patterns of protein binding sites, histone modifications or accessibility data. They represent regions of the genome that have a particular chromatin makeup and, by reducing complex datasets to a set of well-defined functional regions, are a powerful tool to find the relationships between a large number of chromatin proteins.

Several software packages exist to determine and classify the different groups of co-occurrence, one of which is ChromHMM (Ernst and Kellis, 2012). This uses the mapped reads for each dataset to identify binding sites above a Poisson background distribution. From the binding sites,
Distance control of enrichments of interacting fragments overlapping with different modifications over all non-bait fragments, for (a) mESC and (b) fetal liver. Fragments interacting with promoters in the higher expression categories are enriched in marks for active transcription compared with fragments that interact with lowly- or non-expressed promoters.
Figure 5.41: Distance control of enrichments of interacting fragments overlapping with different modifications over all non-bait fragments, for (a) GM12878 and (b) CD34+ cells. Fragments interacting with promoters in the higher expression categories are enriched in marks for active transcription compared with fragments that interact with lowly- or non-expressed promoters.
it then trains a Hidden Markov Model (HMM) to identify a predefined number of states in the datasets provided.

For the GM12878 sample I used the chromatin segmentation profiles created by the ENCODE project using the histone modifications used beforehand and the TFs that were also assayed in this cell type by the same project. The states found were then classified as belonging to promoters, enhancers, insulators, transcription-associated, repressed transcription and repetitive areas of the genome.

For the mESC sample I used a combination of TF and chromatin-binding protein ChIP-seq datasets. The datasets used include, in addition to the histone modifications used in the previous section, the insulator protein CTCF, cohesin subunits Smc1 and Smc3 and Oct4, Nanog, Sox2 and Klf4, which are TFs associated with ESC pluripotency. I then used ChromHMM to identify areas of the genome that present particular combinations of these proteins and the histone marks. The individual contribution of each mark or protein to each of the seven states is provided in Figure 5.43a. States 1 and 3 represent areas bound mainly by the repressive proteins H3K27me3 and H3K9me3, state 2 shows no particular binding, states 4 to 6 represent transcription areas, with activating marks and TFs and finally state 7 is heavily marked by CTCF and the cohesin subunits.

The CD34+ and mouse fetal liver samples have a smaller number of available TF datasets. Therefore, there would be little added information in using the chromatin state approach compared to using just the chromatin marks independently as in the previous sections.
I calculated the enrichment of interacting fragments using the states for the mESC cell type and the GM12878 cell types. The enrichments that the interacting fragments have for the different states were calculated in the same manner as in the previous section for the individual chromatin marks. The results, shown in figures 5.43b and c, are consistent with the results found using the histone modifications alone. Fragments that interact with highly-expressed promoters are enriched in active states and depleted in repressed states in both species and samples.

The results from the analysis of chromatin states support the histone modification analysis and show that non-histone protein data follows
the same pattern. This further reinforces the notion that the interacting fragments reflect the status of the promoters they interact with.

5.2.7 Cohabiting marks in interacting fragments follow known patterns

Having established that the chromatin status of interacting fragments reflect the expression status of the promoters which they contact, I investigated the extent to which these marks are present in the same interacting fragment.

In figure 5.44 I plot the fraction of promoter-interacting fragments that have the both modification listed on the X-axis and the one on the Y-axis, for the mESC and GM12878 samples. The main co-occurring histone marks are those that are associated with transcription. For example, in the mESC sample, the marks that co-occur often with H3K4me3 are H3K9ac and H3K4me2, both associated with transcription. On the other hand, the mark that shows the lowest overlap with any other marks is H3K27me3, which is expected, as this mark showed a different enrichment behaviour compared with the activating marks in the previous analysis.

Figure 5.44: Co-occurrence of histone modifications on interacting fragments in mESC (a) and GM12878 (b) cells. The colours represent the proportion of fragments that have the modification on the X-axis that also have the one on the Y-axis.

Transcription-associated marks have a particular behaviour in their co-occurrence that is in contrast to the behaviour of the mark most associated with repression, H3K27me3. This difference between active marks and H3K27me3 is similar to what was seen before for the enrichments in interacting fragments. Taken together, these results show that the chromatin landscape of interacting fragments reflect the status seen at promoters.
5.2.8 Promoter-promoter interactions

Using our promoter-enriched samples we can not only identify interactions between promoter fragments and all other fragments but also between two promoter fragments, and investigate whether promoter-promoter interactions occur preferentially between promoters in similar chromatin environments.

I used the same set of chromatin modifications used in the promoter-other analysis to query their behaviour in promoter-promoter analysis. I measured how likely a modification $i$ is to be in a promoter interacting with another promoter that has the modification $l$ compared with how likely the modification $i$ is to be seen in any of the promoter-promoter interactions. As with the previous analysis, the results are similar between samples and species. Figure 5.45 shows the results for the mESC and GM12878 samples.

Figure 5.45: Repressive marks H3K27me3 and H3K9me3 are more likely to be present in both sides of a promoter-promoter interactions. The heat maps show the added probability that a modification will occur on both sides of an interaction compared with how likely it is to be present in any interaction, both for mESCs (a) and GM12878 (b). The same results are observed for the mouse fetal liver and human CD34+ samples (not shown).

The size of the effect is quite small, but there is a clear trend for a promoters that has the repressive mark H3K27me3 to interact with another promoter with the same mark instead of with a promoter with any of the other marks assayed. The same is true for promoters with the heterochromatin mark H3K9me3. Marks associated with transcription, conversely, seem to appear together in promoter-promoter interactions. The same pattern seen for interactions between promoter and non-promoters and chromatin modifications is thus seen in promoter-promoter interactions. These trends in promoter-promoter interactions
reinforce the notion of interactions between transcriptionally active promoters and point to the existence of similar interactions between repressed promoters.

5.2.9 **Topologically associated domains (TADs)**

A feature highlighted by most studies that assayed metazoan genomes using Hi-C is the existence of TADs. The higher power of the C-Hi-C technique to detect interactions involving promoters allows us to evaluate the effect of TADs in this subset of the overall interaction landscape.

For every promoter-fragment that falls within an annotated TAD, I calculated the extent to which interactions involving that promoter connect with a fragment that lies further along the chromosome instead of before it. The measure is the simple proportion of fragments that interact with a promoter that have a higher genomic coordinate than the promoter, divided by the total number of fragments that interact with it. This proportion is then scaled to a -1 to 1 interval, where a value of 1 means that all interactions involving that promoter are with fragments with higher genomic coordinates, 0 means that half of the interaction targets have higher genomic coordinates and half have lower genomic coordinates and a value of -1 means that all the interaction targets have lower genomic coordinates. I then identified the position of the bait fragments compared to the genomic co-ordinates of the TADs identified by Dixon et al. (2012) and plotted the histograms for the promoters that fall within each 10% of their TAD. I did this separately for all interactions, for those between promoter and non-promoter fragments and for those involving two promoter fragments.

Figure 5.46: TADs influence interactions between promoter and non-promoter fragments to a greater degree than interactions between two promoters. For every bait fragment inside a TAD I calculated a directionality index that reflects whether most of the interactions are downstream or upstream. The plots show the directionality indexes for all interactions (a), promoter-other interactions (b) and promoter-promoter interactions (c) relative to the positioning of the promoter along the TADs.

If TADs influence the interactions discovered in our samples, we should see an effect when plotting the directionality of interactions involving the promoters along them. Those promoters that fall close to
the beginning of the TAD are expected to interact with downstream fragments, whereas those at the end of the TAD will preferably interact with upstream fragments. This is in fact what I observe for all interactions in figure 5.46a and for interactions involving a promoter and a non-promoter fragment in figure 5.46b. However, when only promoter-promoter interactions are examined, the effect of the TADs is greatly diminished and the trend seen in the other plots is almost non-existent (5.46d). The results shown are for the GM12878 sample, but they are similar in the other three conditions.

TADs are clearly seen in other studies and their effect is present in our datasets for promoter-other interactions, but much less for promoter-promoter interactions. This result raises the question of the validity of TADs for this type of interactions. One possible explanation is that they reflect those proximal interactions that result from folding of the chromosomes, but the more arguably more directed and functional promoter-promoter interactions could be actively set up and so escape the influence of TADs.

5.2.10 CTCF and cohesin

There are many additional non-histone proteins that influence the folding of DNA. Of particular interest is the CCCTC-binding factor, or CTCF, which can block interactions between promoters and enhancers and is described as an insulator. CTCF also localises with the cohesin complex, itself implicated in chromatin conformation. As we have seen, the TADs have a reduced effect in our promoter C-Hi-C datasets. To see how the latter relate to insulators, I used publicly available data to define CTCF and cohesin peaks in mESC and GM12878 cell lines. The raw ChIP-seq data was processed in the same manner as the histone modification datasets.

The first finding is that promoter-other interactions occur across most CTCF and cohesin peaks. In mESC only 2.6% of 33,836 CTCF peaks and 1.4% of 8939 cohesin are not bridged by any interaction. For GM12878 10.7% of 61,625 CTCF peaks and 9.7% of 23,758 cohesin peaks are not bridged. When TADs were first defined, CTCF was found to be present at their borders and one possibility is that the unbridged peaks are those at TAD borders. However, bridged and unbridged CTCF and cohesin peaks show cohesin peaks have a similar distribution of distances to the nearest TAD (5.47a), indicating that it is not the distance of insulator peaks to TADs that determines whether these peaks are bridged or not.

Given that our interaction dataset is genome-wide and includes interactions that can span over a megabase, it is possible that the insulator locations still have an effect and are only bridged by long-range interactions. To see if this is the case, I examined whether the distances spanned by all interactions is different than the distances spanned by
interactions that bridge a CTCF or a cohesin site. The results are presented in figure 5.47b. In that figure, there is an increase in the distance spanned by interactions that bridge cohesin sites (blue line), compared to interactions that bridge CTCF sites (red line) or all interactions (black line). This effect could be reflecting a local effect of insulator proteins, blocking interactions between fragments close to them.

5.2.11 LADs are inert interaction-wise

Lamins are proteins that interact with other proteins in the inner nuclear membrane to form a fibrous network, the nuclear lamina. Previously, lamin-associated domains (LADs) were shown to contain lowly-expressed genes and suggested to form a repressive chromatin environment \cite{Guelen2008}. Using the promoter C-Hi-C datasets, I can determine the extent to which promoters interact with LADs.

To see how the interactions involving promoters behave in relation to LADs, I used a publicly available mouse LAD dataset \cite{Peric-Hupkes2010}. I then separated the promoter-other interactions according to whether the fragments they connect are both outside LADs, both inside the same LAD, both inside two different LADs or one fragment is inside a LAD while the other is outside them.

The results in figure 5.48 show that the vast majority of interactions happens between fragments that are both located outside LADs. In-

---

**Figure 5.47:**

(a) Density plot of distances of bridged and unbridged cohesin peaks to the nearest TAD border. There is no difference between the two groups. (b) Distance spanned by all promoter-other interactions in the mESC sample (black), by the interactions that span CTCF peaks (red) and by the interactions that span cohesin peaks (blue). Local interactions are seemingly underrepresented in the red and blue set, showing local influence of the CTCF and cohesin sites.
terestingly, the next biggest group is those of interactions inside the same LAD, followed by interactions where one only one of the ends is inside a LAD. Interactions between different LADs, on the other hand, are a very small subset of the total interaction number. This means that few promoter-other interactions involve promoters inside TADs and, when they do, these are between two fragments that are inside the same LAD.

![Graph showing interactions of mESC sample](image)

Figure 5.48: Most interactions of the mESC sample happen outside LADs. Interestingly, the second biggest group is that of interactions between two fragments inside LADs.

### 5.3 Discussion

In the work presented here I characterised the interactions involving promoters of *M. musculus* and *H. sapiens*. The technique used, promoter C-Hi-C, provides a balance between the resolution of 4C and the genome-wide properties of Hi-C. Potentially all interactions involving a promoter can be identified.

That the chromatin makeup of a promoter region is associated with the expression status of that promoter is well established. Here I show that non-promoter regions that interact with promoters also reflect the expression status of those promoters. Long-distance regulatory regions like enhancers are known, but my results indicate that areas not previously characterised as enhancers also present a different chromatin makeup depending on the activity of the interacting promoter. One possibility is that the number of long-distance regulatory regions is higher than previously known. These "low-profile" enhancers might have smaller levels of the characteristic histone modifications and thus
have not been detected using ChIP-seq. Another possibility is that the chromatin makeup of these promoter-interacting regions is not a cause, but a consequence of the promoter status. If that is the case, the physical proximity of the promoter to the interacting fragment lead to the enzymes that deposit these marks on the former depositing them also on the latter.

The observation regarding promoter-promoter interactions and their chromatin marks is particularly interesting. Having promoters with active marks interacting with promoters with active marks and promoters with repressive marks interacting with promoters with repressive marks reinforces not only the idea of compartments in the nucleus where the transcription of several promoters co-occurs (e.g. transcription factories) but also suggests that there are analogous compartments that turn off co-regulated genes.

The results from this analysis point out that highly-expressed promoters seem to interact preferentially with other regions bearing hallmark marks of gene expression. One possibility is that the promoters interact mainly with their proximal regions and genes located closely to each other tend to have similar expression levels (Spellman and Rubin, 2002). Supporting this possibility is the fact that other chromatin interactions studies have observed that genes that belong to the same TAD also tend to be at similar activity levels (Sexton et al., 2012). However, the pattern of increased enrichment of active chromatin marks in the fragments interacting with highly expressed promoters compared to lowly expressed promoters is observed even when very long range interactions are taken into account, suggesting that this observation is not only due to the gene neighbourhood.

Being placed on the middle of the 3C-technique scale between the extensiveness of the area assayed and the sensitivity of the assay, promoter C-Hi-C allows for a deeper understanding of the overarching chromatin architecture identified by Hi-C. The fact that TADs have a greater influence on interactions between promoters and non-promoters than on interactions between promoters leads to the question of whether the TADs are caused by a functional preference of interactions or by just by the packaging of DNA and the physical proximity of two loci. We can envision a scenario in which two condensed TADs have biologically important interactions (such as those between two co-regulated promoters), reaching out of the domains and to each other. Even if that is not the case, the panorama of interactions inside a nucleus is more complex than discrete domains of interactions. The repressive lamin-associated domains, on the other hand, do seem to constrain interactions involving promoters.

The results found using these first promoter C-Hi-C datasets call attention to the extent to which the different techniques used to capture chromosome conformation illuminate different aspects of the complexity inside the nucleus. In due time, perhaps it will be possible and useful
to combine the different types of experiments into a meta-interactome that can explain more than the sum of its parts.
<table>
<thead>
<tr>
<th>Histone modification</th>
<th>GEO accession ID</th>
<th>Publication</th>
</tr>
</thead>
<tbody>
<tr>
<td>H3K4me1</td>
<td>GSM723016</td>
<td>Shen et al. (2012)</td>
</tr>
<tr>
<td>H3K4me3</td>
<td>GSM723017</td>
<td>Shen et al. (2012)</td>
</tr>
<tr>
<td>H3K27ac</td>
<td>GSM851278</td>
<td>Shen et al. (2012)</td>
</tr>
<tr>
<td>H3K27me3</td>
<td>GSM1000089</td>
<td>Mouse ENCODE Consortium et al. (2012)</td>
</tr>
<tr>
<td>H3K36me3</td>
<td>GSM1000109</td>
<td>Mouse ENCODE Consortium et al. (2012)</td>
</tr>
<tr>
<td>H3K9ac</td>
<td>GSM1000127</td>
<td>Mouse ENCODE Consortium et al. (2012)</td>
</tr>
<tr>
<td>H3K9me3</td>
<td>GSM1000147</td>
<td>Mouse ENCODE Consortium et al. (2012)</td>
</tr>
<tr>
<td>H3K4me2</td>
<td>GSM747543-GSM747544</td>
<td>Stadler et al. (2011)</td>
</tr>
</tbody>
</table>

Table 17: Histone modification datasets used in mouse embryonic stem cells.
<table>
<thead>
<tr>
<th>Histone modification</th>
<th>GEO accession ID</th>
<th>Publication</th>
</tr>
</thead>
<tbody>
<tr>
<td>H3K4me1</td>
<td>GSM851299</td>
<td>Shen et al. (2012)</td>
</tr>
<tr>
<td>H3K4me3</td>
<td>GSM851300</td>
<td>Shen et al. (2012)</td>
</tr>
<tr>
<td>H3K27ac</td>
<td>GSM851302</td>
<td>Shen et al. (2012)</td>
</tr>
<tr>
<td>H3K4me2</td>
<td>GSM688808</td>
<td>Wong et al. (2011)</td>
</tr>
<tr>
<td>H3K9ac</td>
<td>GSM688810</td>
<td>Wong et al. (2011)</td>
</tr>
<tr>
<td>H3K36me3</td>
<td>GSM688812</td>
<td>Wong et al. (2011)</td>
</tr>
<tr>
<td>H3K79me2</td>
<td>GSM688813</td>
<td>Wong et al. (2011)</td>
</tr>
<tr>
<td>H4K16ac</td>
<td>GSM688814</td>
<td>Wong et al. (2011)</td>
</tr>
<tr>
<td>H3K27me3</td>
<td>GSM688811</td>
<td>Wong et al. (2011)</td>
</tr>
<tr>
<td>H3K9me3</td>
<td>GSM946549</td>
<td>Mouse ENCODE Consortium et al. (2012)</td>
</tr>
</tbody>
</table>

Table 18: Histone modification datasets used in the analysis of mouse fetal liver cells.
Histone modification | GEO accession ID | Publication
--- | --- | ---
H2A.Z | GSM733767 | ENCODE Project Consortium et al. (2012)
H3K27ac | GSM733771 | ENCODE Project Consortium et al. (2012)
H3K27me3 | GSM733758 | ENCODE Project Consortium et al. (2012)
H3K36me3 | GSM733679 | ENCODE Project Consortium et al. (2012)
H3K4me1 | GSM733772 | ENCODE Project Consortium et al. (2012)
H3K4me3 | GSM733708 | ENCODE Project Consortium et al. (2012)
H3K79me2 | GSM733736 | ENCODE Project Consortium et al. (2012)
H3K9ac | GSM733677 | ENCODE Project Consortium et al. (2012)
H3K9me3 | GSM733664 | ENCODE Project Consortium et al. (2012)
H4K20me1 | GSM733642 | ENCODE Project Consortium et al. (2012)

Table 19: Histone modification datasets used in the analysis of GM12878 cells.

Histone modification | GEO accession ID | Publication
--- | --- | ---
H3K4me1 | GSM706845 | Bernstein et al. (2010)
H3K4me3 | GSM773041 | Bernstein et al. (2010)
H3K9me3 | GSM772938 | Bernstein et al. (2010)
H3K27me3 | GSM772951 | Bernstein et al. (2010)
H3K27ac | GSM772870 | Bernstein et al. (2010)
H3K36me3 | GSM772865 | Bernstein et al. (2010)

Table 20: Histone modification datasets used in the analysis of CD34+ cells.
6.1 SUMMARY

In this thesis I explored several aspects of the relationship between chromatin and the regulation of transcription. In chapter 1, I described in general terms eukaryotic transcription and how chromatin modifications and chromatin structure can influence the transcriptional process. In chapter 2, I investigated the behaviour of Mof and Msl1, two proteins involved in the up-regulation of genes on the male X chromosome of Drosophila melanogaster as part of the dosage compensation mechanism. Chapter 3 continued the focus on Drosophila dosage compensation, but from the perspective of roX1 and roX2, the ncRNA components of the dosage compensation complex (DCC), and of Mle and Msl2, the proteins that interact with the roX RNAs. The project described in chapter 4 used publicly available data to describe different modes of binding of RNA Polymerase II in D. melanogaster promoters. In chapter 5, I investigated the connection between DNA interactions involving promoters and the chromatin content of the interacting regions.

In this final chapter, I will briefly describe the main results from each of these projects. I will also expound on the current state of computational work in the field of genomics, its relationship with the wet lab and the impact that data-producing consortia and data repositories have in molecular biology research.

6.2 CONSERVED FUNCTIONS OF MOF AND MSL1 IN DROSOPHILA

Dosage compensation in D. melanogaster is a model of up-regulation of gene expression on a genomic scale. To compensate for the difference in the number of X chromosomes in male and female flies, the sole male X chromosome has its expression increased two-fold. The proteins involved in dosage compensation in fly are conserved in vertebrates, where dosage compensation is achieved through a different mechanism and by different proteins. This suggests that dosage compensation proteins in fly have other functions that may be common to metazoans. As part of a collaboration between the Luscombe Laboratory of the London Research Institute and the Akhtar Laboratory of the Max Planck Institute for Immunobiology and Epigenetics, I characterised the evolutionary conserved behaviour in two fly species of Mof and Msl1, two proteins that are part of the fly DCC.
For the evolutionary comparison, I identified the binding locations of Mof and Msl1 in two Drosophila species that diverged around 40 million years ago: D. melanogaster and D. virilis. Mof binds along the body of X-linked genes in both species. Both proteins also bind promoters in both X-linked and autosomal genes in the two species. I established that the sequence preference at high affinity sites for the DCC is a similar GA-rich motif in both species. Furthermore, by evaluating the conservation of binding in sets of orthologous genes, I determined that the gene body binding of Mof is dependent on the gene being located in the X chromosome. By contrast, the promoter-binding mode of Mof is conserved regardless of the chromosome where the genes are located.

While the genome-wide binding of Mof at promoters has already been characterised in previous studies, less is known about the similar behaviour of Msl1. To determine the extent to which Msl1 binding at promoters is male- or Mof-dependent, I identified its binding locations in female samples and in male samples where Mof has been truncated and rendered non-functional by the loss of its chromodomain and of its histone acetyl-transferase domain. Msl1 also binds promoters in these two conditions and is linked to higher expression levels in both wild-type and Mof mutant male samples. Thus, Msl1 has a promoter-binding function linked to gene expression that acts genome-wide in a manner independent of Mof and dosage compensation.

We next attempted to elucidate the mechanism behind the Msl1 link to expression. After Msl1 RNAi, we observed a depletion in the global levels of RNA polymerase II serine 5 phosphorylation. Msl1 depletion also led to a depletion at promoters of CDK7, the kinase responsible for this phosphorylation, but not to a diminution of its overall level in the cell. Conversely, in CDK7-inhibiting conditions Msl1 localisation is disrupted. We hypothesise that Msl1 cooperates with CDK7 to carry out properly the phosphorylation of serine 5 of the CTD necessary for transcription initiation.

6.3 Protein-RNA Interactions in Fly Dosage Compensation

Also in collaboration with the Akhtar Laboratory, I looked into the interactions between two other proteins that are part of the DCC, Mle and Msl2, and the ncRNAs that participate in dosage compensation in fly, roX1 and roX2. While the interactions between the different protein components have been characterised via their structure, less is known about how the roX RNAs interact with protein portion of the DCC.

Using individual nucleotide resolution UV cross-linking and immunoprecipitation (iCLIP), we identified the transcripts to which Mle and Msl2 bind. The two transcripts that show the majority of binding events are roX1 and roX2. In roX1, the two proteins bind to three sep-
arate regions, with the highest number of binding events occurring at the 3' of the transcript, which is homologous to roX2. In roX2, the two proteins bind a single region which corresponds to the 3' bound region of roX1. The main peaks of Mle binding in the two roX RNAs localise to conserved sequence motifs that are essential for male survival, the roX boxes. The peaks of Mle binding in roX2 also correspond to regions that are predicted to form double-stranded secondary structures.

To confirm the binding of Mle to roX's secondary structures, we used 2'-hydroxyl acylation analysed by primer extension (SHAPE). This technique identifies paired nucleotides and allows for the determination of the secondary structure of RNAs. Indeed, we confirmed that the peaks of Mle binding and the roX boxes occur in regions that form double-stranded structures. A genetic analysis of the loops detected in roX2 revealed that they are redundant, in that any single disruption would not reduce male viability, but the combination of disruption of several loops gravely hindered male survival.

Using a combination of genomic and structural techniques, we were able to characterise the binding of Mle to double-stranded conserved regions of roX RNAs. The two RNAs are redundant in their function, but we also detect further redundancy inside each RNA, in the stem-loop structures that they require for integration with the DCC. Given the essential nature of dosage compensation to male fly survival, the two layers of redundancy in roX RNAs contribute to the robustness of the fly system as a whole.

6.4 DISTINCT MODES OF POL II BINDING IN FLY

The architecture of chromatin and the DNA sequence at metazoan promoters have been used to identify different types of promoters. Using publicly available datasets of RNA Pol II occupancy in *D. melanogaster* larvae and S2 cells, I identified five distinct modes of binding of this complex at transcription start sites. I combined the five modes with data relating to gene expression, core promoter motifs, transcription factor binding sites, nucleosome positioning, histone post-translational modifications, chromatin status and Gene Ontology terms to try to identify similarities and differences between the different binding modes and explain their biological impact.

Two of the binding modes show a pronounced peak of Pol II signal at the TSS and show a higher proportion of expressed genes than the other modes. One of them has a broader peak and is enriched for genes related to house-keeping functions and for the associated core promoter motifs. The other, with a sharper peak of Pol II signal, is enriched for binding sites of developmental TFs and for genes with developmental functions and their associated core promoter motifs.

Two other modes of Pol II binding are characterised by a peak of Pol II signal that is shifted upstream compared with all other modes. They
show the highest proportion of stalled genes, although even expressed
genes in these modes have the characteristic upstream shift of Pol II. It
is not clear what class of genes or promoters these two modes represent.
A fifth mode of binding shows no pronounced peak of Pol II signal and
represents mainly unexpressed genes.

A factor that is constant across the different modes of Pol II binding
is nucleosome positioning around promoters. This suggests that they are
not related to the primary structure of chromatin. More work will
need to be carried out to characterise fully the observed groups of
transcription start sites. Each Pol II shape includes a mix of classes:
while some modes are enriched in expressed or stalled, developmental
or house-keeping genes, there is not an absolute separation. Perhaps
further, clearer divisions are to be found by incorporating data about
basal transcriptional machinery and chromatin remodelers that prime
promoters to assemble the Pol II holoenzyme.

6.5 C-HIC IDENTIFIES PROMOTER-INTERACTING REGIONS

Capture HiC (C-HiC), a new technique developed by the Fraser and
Cameron laboratories of the Babraham Institute in Cambridge, intro-
duces a library enrichment step to HiC protocols. This modification can
overcome HiC’s resolution limitations linked to sequencing depth, by
focusing the subset of interactions evaluated. In collaboration with the
two laboratories, we analysed C-HiC datasets that capture interactions
involving gene promoters in human and mouse.

Using the map of promoter interactions I established that the pro-
moters of genes with higher expression interact with greater numbers
of enhancers. Furthermore, the non-promoter regions interacting with
highly expressed promoters are enriched for histone marks associated
with active chromatin, while the regions that interact with lowly ex-
pressed promoters show a depletion for the same marks. This applies
to all non-promoter regions, even those that do not contain known en-
hancers, which are marked by a specific set of histone modifications.
The regions that interact with highly-expressed promoters are also en-
riched for states showing a combination of histone marks, TFs and other
chromatin proteins that reflect transcriptionally active chromatin and
are depleted for other states that mark for repressive or inactive chro-
matin. With these results, I show that the chromatin states at distal
regions of the genome reflects the transcriptional activity of the pro-
moters they interact with.

In a similar analysis using the set of promoter-promoter interactions,
I show that promoters with histone marks associated with transcription
tend to interact with promoters that also contain transcription-
associated marks. Similarly, promoters with histone marks associated
with transcription repression such as H3K27me3 and H3K9me3 tend to
interact more with other promoters with the same mark, and less with
promoters with active marks. The concept of transcription factories, nuclear locations where several distal genes are jointly transcribed, has been proposed, and my results suggest that there are also long-distance contacts of repressed genes that may form “shut-down” transcription factories.

The nature of the C-HiC experiments allows me to observe the effect that known genomic features have on interactions involving promoters. One such type of feature in particular is of interest, as it was discovered in the first genome-wide interaction maps: topologically-associated domains (TADs). These domains define highly interconnected stretches of the genome that have fewer interactions with areas outside their own domain. In our C-HiC datasets, TAD boundaries have different levels of impact on interactions involving promoters depending on whether the interaction is between a promoter and a distal region or between two promoters. The former are more affected by the TAD borders, while the latter are less affected and are more likely to cross borders, suggesting that promoter-promoter interactions “leap-out” of established domains.

The results of the promoter C-HiC experiments shed light on the nature of the genomic regions that interact with promoters and into how the subset of promoter-involving interactions is projected on the overarching chromatin make-up and three-dimensional genome architecture. I demonstrate an association between a promoter’s expression level and the chromatin state of its interacting regions.

6.6 COMPUTATIONAL BIOLOGY IN TODAY’S GENOMICS

In addition to the regulation of transcription via chromatin, another thread runs through the different chapters of my thesis: the use of computational techniques for data processing, analysis and visualisation. In this last section, I will attempt to describe my view of the importance of computational work in today’s biology and on how it affects the relationship between the dry computational and the traditional wet laboratories.

6.6.1 The importance of computational work in data analysis

Computational work is not an essential part of every facet of biology yet. However, in some specific areas of biology, no work is complete without a considerable amount of computational work to make sense of the data gathered by the experimentalists. In genomics, computational work is not just another tool in the researcher’s bag, it is an essential part of any project.

The reasons behind the centrality of computational work in genomics are the amount and types of data that this field utilises. Even the simplest eukaryote model organism, Saccharomyces cerevisiae, has 6692
protein coding genes, spread over 12,157,105 base pairs of its genome. To do an analysis of a single variable on the gene level alone already necessitates thousands of data points. The amount of data increases dramatically with the experimental techniques used. As described in chapter 1, several of the techniques used today to perform genome-wide assays use massively parallel short-read sequencing as the readout. Currently, a single ChIP-seq experiment will yield tens of millions of short DNA sequences. Such a technique would not be feasible — or indeed developed — without the assumption that there is specialised software available to map these reads and, later, to convert the mapped locations into meaningful biological features like protein binding sites.

6.6.2 Publicly available data and its impact

In chapter 1 I alluded briefly to experimental data repositories and to large data-generating consortia that make genomic data available to the whole community. Two of the chapters in this thesis illustrate how this data can be used in research. Chapter 4 describes a project that uses solely publicly available data. Chapter 5 also uses publicly available data, but rather than using it to create a whole new project, this data is used to deepen the analysis done using a newly-created dataset. In both cases, datasets generated for a particular project were, after publication, repurposed to answer new biological questions. The tools used to process and analyse the data are also publicly available for research purposes. As I did it, so can anyone else with an internet connection and the appropriate biological and computational knowledge.

The fast expansion of available datasets will only increase the trend of using publicly available data, alone or in conjunction with new experimental results. Despite the many datasets already available for most model organisms, there is still scope for improvement, especially in the number of conditions assayed, i.e., in a wider variety of tissues, developmental stages or cell lines. This expansion need not happen all in a coordinated fashion in large consortia. It will naturally happen as individual laboratories need the missing experiments in their projects and fill in the gaps, perhaps drawing from the experience of the large consortia, who honed and standardised the experimental protocols for big-scale projects.

6.6.3 Wet-dry crosstalk

If there is such a wealth of published datasets ripe for the taking of computational biologists, is there still a role for the traditional wet lab in genomics, other than data generation? The answer, in my opinion, is a resounding yes. Most genomic techniques require biological material from whole populations of cells. This limits their power to observing a snapshot of a population of cells. As exemplified in chapters 2 and 3
of this thesis, “classic”, non-genomic molecular biology techniques are essential to fully elucidate the mechanisms by which we can explain the genome-wide observations of a population of cells.

As the field currently stands, most computational biologists are, by education, biologists. Their computer programming skills are, in general, self-taught or learnt in short courses during their post-graduate career. While this limits the complexity and sophistication of the scripts and software developed in most computational biology laboratories — laboratories fully dedicated to developing new algorithms are a clear exception — their biological background allows them to communicate with their colleagues from the wet lab. Still, as the transition to computational work tends to happen early in the career and the nature of this work facilitates changes between different model organisms and systems, dry biologists benefit from the more in-depth biological knowledge that their wet laboratory colleagues usually have of specific biological systems. In the future, the dissemination of genome-wide techniques will lead to a higher number of hybrid laboratories, with wet and dry biologists working closer together.
BIBLIOGRAPHY


germa: Background Adjustment Using Sequence Information.


data.table: Extension of data.frame for fast indexing, fast ordered joins, fast assignment, fast grouping and list columns.


Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., van Calcar, S., Qu, C., Ching, K. A.,


*R: A Language and Environment for Statistical Computing.*


A.1 General Software Used

Unless otherwise specified, all scripts used to process and visualise the data used throughout this thesis were done using the R programming language (rci). Manipulation of genomic intervals, coordinates and overlaps was done using the GenomeRanges (Abovoun et al., 2013) package from the Bioconductor project (Gentleman et al., 2004). Plots were created using the basic R plot functions and the ggplot2 package (Wickham, 2009). The data.table package (gen) was used for efficient computation of values in subsets of large data frames.

A.2 Methods for Chapter 2

A.2.1 Drosophila ChIP-seq processing

The results of all ChIP-seq experiments for MOF and MSL1 in D. melanogaster and D. virilis were processed in an identical way.

Read alignment

Reads from both IP and input samples were mapped to the reference D. melanogaster or D. virilis genome using Bowtie (Langmead et al., 2009). Default parameters were used, which allows for 2 mismatches in a seed 28 nucleotides long. The only deviation from default parameters was the “best” option, which stipulates that when a read aligns to several genomic locations only the highest scoring one is reported. The command line used was as follows:

bowtie --best <genome_reference_file> <sequenced_reads_file> <output_file>

Determining significantly-bound regions

Significantly-bound regions were determined using an approach developed by Florence Cavalli and Juanma Vaquerizas in Conrad et al. (2012b). The approach was implemented using R scripts (rci).

The reference genome was divided into bins with a width of 25 bp and the number of reads from the IP or input samples that overlapped with each bin was counted. DESeq (Anders and Huber, 2010) was then used to normalise the number of reads of each sample in each bin using a factor derived from the sample size. The resulting IP normalised count was then divided by the normalised input count and this ratio
was converted to its log2 value. The log2 fold-change values of each chromosome were smoothed by taking the mean of a 400 bp-wide sliding window.

All real binding sites are assumed to be present only in a long right-hand tail distribution of the smoothed log2 fold-values. Therefore, a symmetric null distribution was created by joining the distribution of values to the left of the mode with its mirror image. Using the null distribution for each ChIP-seq experiment a FDR-adjusted p-value cut-off of 0.05 was identified. All genome bins below this cut-off were considered significantly bound.

Creating average gene profiles

In order to create average gene profiles of the ChIP-seq signal the coordinates of *D. melanogaster* and *D. virilis* were downloaded from Ensembl ([Flicek et al., 2012](#)). The smoothed log2 fold-change values of bins that fell within a 4000 bp window centred on the TSS, defined by the gene start coordinates in Ensembl, were taken and the mean value for each position was calculated. A similar average was calculated for the set of bins which overlapped with the interval 2000 bp upstream of the TTS and the TTS, as determined by the gene end coordinate of Ensembl.

For each sample, average gene profiles were constructed using the set of all genes, autosomal genes and X-linked genes. Within these categories, a further two averages were calculated using the subset of genes that were considered bound by MOF or MSL1 at the promoter and those that were not.

Determining bound genes

Given the existence of a peak of ChIP-seq signal upstream of the TSS in both species, a window of 200 bp centred on the average peak was used as the area to determine genes bound at the promoter. For *D. melanogaster* this window extended from 200 bp upstream of the TSS to the TSS, while for *D. virilis* the window starts 375 and ends 175 bp upstream of the TSS. Any gene for which there is a significantly-bound bin within this window is classified as being bound at the promoter by the protein.

To determine which genes are bound along the gene body the number of bins bound in its body were assayed. If more than 60% of a gene’s bins that overlapped its exons were significantly-bound, the gene was classified as being bound in the gene body. Genes that were less than 600 bp long were excluded from the gene body binding classification.
A.2.2 Assaying binding conservation

Conservation of binding in orthologous genes

A list of one-to-one orthologous pairs of genes of *D. melanogaster* and *D. virilis* was downloaded from the Ensembl Compara database (Kersey *et al.*, 2012). An heatmap of orthologs was constructed by plotting the smoothed log2 fold-change values of the bins that fall under the same windows as the average gene plots, but individually for each gene. Orthologous gene pairs were plotted side by side.

To access the extent to which binding by MOF was dependent on the chromosome on which the genes are located, two subsets of ortholog pairs were selected: those where both the *D. melanogaster* and the *D. virilis* genes are located on the respective X chromosome and those where one of the members of the pair is located on the X chromosome and the other is located on an autosome. For each of these subsets the proportion of pairs that are classified as being bound by MOF in both species, only in *D. melanogaster* or *D. virilis* was calculated. This calculation was done for MOF binding at the gene body and at the promoter.

Conservation of HAS sequence specificity

To identify putative HAS, the regions with highest MOF and MSL1 signal in each species were used. TO this end, those bins in the top 0.1% percentile of smoothed log2 fold-change value for MOF were selected, as well as those in the top 0.1% percentile of smoothed log2 fold-change value for MSL1 and those bins that belonged to both categories were used as putative HAS. The underlying sequence was surveyed for sequence motifs using MEME (Bailey *et al.*, 2006). An identical approach was used for the *D. virilis* HAS using that species’ MOF and MSL1 samples.

A.2.3 Expression analysis

GeneChIP array processing

The Affymetrix GeneChip arrays triplicates for wild type and mutant larvae were normalised using the GCRMA Bioconductor package. The assignments between array probesets and *D. melanogaster* genes were downloaded from Ensembl. The mean of all probesets that assigned to a gene was used as the expression value of that gene.

Expression plotting

Genes were separated by their MSL1 binding status and their genomic location (autosomal or X-linked) and their expression values were plotted using box plots. A further separation was done based on whether
the promoter of the gene showed any presence of H4K16 acetylation in the respective cell type. The H4K16ac data was provided by Florence Cavalli of the Luscombe Lab based on ChIP-seq experiments done using third instar wild-type and mof2 larvae.

A.3 METHODS FOR CHAPTER 3

A.3.1 MLE ChIP-seq

Read processing and peak calling

The processing of the D. melanogaster MLE ChIP-seq experiment was done in a manner identical to the MOF and MSL1 samples in chapter 1. For full details of the processing see section A.2.1.

Comparison of MLE binding and HAS location

In order to compare the binding of MLE to the location of HAS, the union of two sets of previously published HAS (Alekseyenko et al., 2008; Straub et al., 2008) was taken. The genome browser comparing the location of HAS and MOF and MLE binding was produced using the Integrated Genome Browser (Nicol et al., 2009). For the visualisation of the overlap of MLE and HAS in the X chromosome, genes whose Ensembl coordinates overlap with a MLE-bound genome bin were identified. Those MLE-bound genes that also overlapped with a HAS were then plotted according to their coordinates and strand on the X chromosome.

A.3.2 MLE and MSL2 iCLIP processing

Read alignment

Reads from each of the triplicates of the MLE and MSL1 iCLIP experiments were aligned to the reference D. melanogaster genome using BWA (Li and Durbin, 2009). BWA parameters were set so that the 5 nt of the random barcode would not be aligned, a seed of 30 nt was used for mapping and no mismatches were allowed. Specifically, the options used were:

```
bwa aln -n 0 -l 30 -k 0 -B 5 <genome.fasta> <reads.fastq> \<output.sai>
```

The output sai files were then converted to the SAM format.

Determining significantly-bound nucleotides

The approach to identify which nucleotides are significantly bound in the iCLIP experiments was based on the one taken by König et al.
The pipeline was implemented using R scripts. For each individual replicate, the number of reads with non-identical barcodes that map to each nucleotide in the genome is counted. This is taken to be the number of individual binding events detected at those nucleotides. Binding events that occurred within 30 nt of each other were clustered together.

Using the observed number of binding events for each exon or intron in the genome, a thousand random permutations of binding events were created. A FDR of 0.05 was estimated for each of the exons or introns from the random permutations, which was then used as a threshold for selection of significant binding events. The random permutations were done separately for each replicate. Those nucleotides that pass the threshold in two or more of the replicates are considered significantly-bound and the lowest number of random barcodes among the replicates was assigned as the score for those nucleotides.

**roX2 secondary structure prediction**

The secondary structure of roX2 was predicted by extracting the sequence of the roX2 transcript from Ensembl and inputing it into the RNAfold software (Hofacker et al., 1994).

### A.4 Methods for Chapter 4

#### A.4.1 Definition of Pol II clusters

**Pol II ChIP-seq**

Previously published Pol II ChIP-seq datasets from male and female third instar larvae salivary glands and S2 cells were used to assign promoters to Pol II clusters (Conrad et al., 2012b). Florence Cavalli of the Luscombe Lab provided the log2 fold-change values of IP over input for 25 bp-wide bins along the entire genome. The processing steps from raw reads to log2 fold-change values were identical to those described in section A.2.1.

**Expression arrays**

Affymetrix GeneChip arrays from the same study (Conrad et al., 2012b) and in the same conditions as the ChIP-seq Pol II samples were used to assay gene expression. Triplicates were available for each condition (male and female larval salivary glands and S2 cells).

All processing of the expression arrays was done using R scripts and Bioconductor packages (Gentleman et al., 2004). Normalisation between replicates and samples was done using GCRMA (gcr). Expression calls for individual probesets were obtained via the implementation of the Affymetrix MAS5 algorithm in the affy Bioconductor package (Gautier et al., 2004). Probesets with an interquartile range of more
than 0.5 were tested for differential expression between the male and female samples using limma (Smyth, 2005).

**Definition of TSSs to be clustered**

The coordinates of all unique *D. melanogaster* TSSs were downloaded from Ensembl. Short capped RNA sequencing data (Nechaev *et al.*, 2010) was used to adjust the Ensembl TSS coordinates, using the new coordinates provided for existing Ensembl transcripts IDs in that study’s supplementary materials.

All TSSs from the X chromosome were excluded. TSSs from genes that were classified as differentially expressed between the Affymetrix GeneChip arrays from male and female samples were also excluded from the analysis.

**Clustering of TSSs based on the Pol II profile in the male and female samples**

A window was defined extending from 300 bp upstream to 700 bp downstream of all TSSs to be analysed. The log2 fold-change values of the 25 bp-wide bins that fell within these windows were converted to their derivative. For each bin, the derivative was taken as the ratio of the log2 fold-change values of the immediately downstream bin over the immediately upstream bin. The vectors of the Pol II signal derivative were clustered using k-means clustering.

The resulting clusters were visualised by plotting the log2 fold-change values and their corresponding derivative in heat maps where the vectors corresponding to the window around each TSS were sorted by the cluster they belonged to. Average profiles were constructed by averaging the log2 fold-change values of the TSSs of each cluster at individual positions in the window around the TSS. In addition to the average Pol II profiles of each cluster, separate profiles were constructed for the sets of TSSs belonging to genes that were located more or less than 500 bp downstream of another gene and for the sets of TSSs belonging to genes that were in the same or in the opposite orientation to the upstream gene.

**Clustering of TSS in S2 cells**

The Pol II ChIP-seq log2 fold-change values from S2 cells were taken from the same window between 300 bp upstream and 700 bp downstream of the TSSs and converted to their derivate. The euclidean distance between each of these derivative vectors and the derivate of the average profile of the five clusters was calculated. Each S2 TSS was then assigned to the cluster to which its derivative vector showed the smallest distance. The average Pol II profiles for the clusters in the S2 cell type were constructed in the same manner as the average Pol II profiles for the male and female salivary gland samples.
A.4.2 Pol II clusters and expression status

Definition of expression status

In each sample, transcripts were assigned an expression status based on the expression array and Pol II ChIP-seq data. Transcripts were classified as expressed if all the probesets that mapped to it received positive MAS5 calls, if the Pol II log2 fold-change value at its TSS was above the significance threshold for that sample and if the median value of Pol II log2 fold-change values in its gene body was positive. Stalled transcripts received negative MAS5 calls in all their probesets, had a negative median value of Pol II log2 fold-change values in their gene bodies and a Pol II log2 fold-change value at their TSSs above the significance threshold for that sample. Average gene profiles were constructed for the sets of expressed and stalled transcripts inside each cluster.

Expression across the developmental program

For measures of the expression levels of transcripts across the fly developmental program the FPKM values for each transcript from RNA-seq experiments done in 30 different developmental conditions were downloaded from one of the modENCODE studies (Graveley et al., 2010). The interquartile range (IQR) of FPKM values of each transcript in the 30 conditions was calculated and the resulting IQRs of the transcripts in each cluster were plotted as notched box plots.

A.4.3 Pol II clusters and core promoter elements

The locations of predictions by the McPromoter software (Ohler, 2006) based on the presence of core promoter elements in the underlying DNA sequence were downloaded. If one of the clustered TSSs was located within 50 bp of a McPromoter prediction site it was assigned the same core promoter element model as the prediction. The proportion of members of each cluster that was assigned to each core promoter element model was then calculated.

A.4.4 Pol II clusters and TF binding

Binding sites for the Caudal, Ecdysone receptor, Deformed and Nejire transcription factors were downloaded from the modENCODE consortium website as coordinates of genomic intervals that showed a significant signal from ChIP experiments in third instar larvae. Each TSS was subsequently classified as being bound by one of the TFs if a window of 400 bp around the TSS overlapped in any coordinate with one of the TF binding site genomic intervals.
The distance between a TSS and an overlapping TF binding site was calculated as the difference in base pairs between the TSS coordinate and the middle point of the TF binding site. Using the numbers of TSSs that overlapped with a TF binding site, the enrichments of each TF in each Pol II cluster were calculated and subsequently visualised in the form of a heatmap. The enrichment calculation was done using the following equation:

$$enrichment = \log_2 \left( \frac{\#\text{TSSsInClusterWithTF}}{\#\text{TSSsInCluster}} \frac{\#\text{TSSsWithTF}}{\#\text{TSSs}} \right)$$

(A.4.5) Nucleosome analysis

Defining nucleosome characteristics

Paired-end reads from an MNase-seq experiment in S2 cells (Gilchrist et al., 2010) were mapped to the reference D. melanogaster genome using Bowtie (Langmead et al., 2009) in default parameters. The location of the mapped reads was converted by GeneTrack (Albert et al., 2008) into nucleosome positions.

For each TSS the +1 nucleosome was either the nucleosome whose occupancy coordinates overlapped with the TSS or, should there be no nucleosome overlapping the TSS, the first nucleosome downstream of the TSS. The -1 nucleosome was defined as being the first nucleosome upstream of the +1 nucleosome. The Nucleosome Free Region (NFR) was measured as the interval that lies between the end of the coordinates occupied by the -1 nucleosome and the start of the coordinates occupied by the +1 nucleosome. Finally, the focus of each nucleosome was measured as the ratio between the height of the GeneTrack peak for that nucleosome and the total number mapped reads that show any overlap with the nucleosome occupancy coordinates.

Visualising nucleosome data

In order to create the plots showing the variation of nucleosome characteristics (NFR, +1 and -1 nucleosome focus) as gene expression changes, the TSSs were sorted according to their microarray expression level. A sliding window of 100 of these sorted TSSs was then taken, with the minimum, first quartile, median, third quartile and maximum value of each of the sliding windows being plotted.

Average nucleosome profiles were created by taking a window of 4000 bp around each TSS and measuring the proportion of base pairs in each position that were occluded (that is, that overlapped with a nucleosome position from GeneTrack). Individual average nucleosome profiles were created for each Pol II cluster. Further subsets were used and the respective average profiles were created based on the Expressed/Stalled expression classification based on the microarray and Pol II ChIP-seq
data and based solely on the microarray data MAS5 calls to define a transcript as being expressed or not.

**Histone modifications and variants**

The positions of post-translationally modified nucleosomes or nucleosome variants in S2 cells, provided as coordinates of regions that showed significant signal from ChIP experiments were downloaded from the modENCODE consortium website (Kharchenko et al., 2010). A specific nucleosome modification or variant was considered as being present in a TSS if there was any overlap between the genomic interval with significant signal for the modification or variant and a window of 400 bp around the TSS.

The enrichment of individual histone modifications or variants among the TSSs of the Pol II clusters was calculated using the equation 3, the same used for calculating the enrichment of TF binding sites. Further enrichments for histone modifications or variants among the subset of each cluster’s TSSs that belonged to expressed or not expressed transcripts were also calculated using equation 3. All enrichments were visualised in heatmap form.

**Pol II clusters and chromatin states**

A segmentation of the *D. melanogaster* genome into distinct chromatin states was done by Filion et al. (2010) according to the binding of chromatin proteins as assayed by DamID. The resulting segmentation annotation was downloaded as the genomic intervals that correspond to each of the chromatin states. Each TSS was assigned to one of the chromatin states according to what was the chromatin state in the exact coordinate of the TSS. The number of TSSs that correspond to each combination of Pol II cluster, chromatin state and expression status were counted and the results visualised using a stacked bar plot.

**GO analysis**

Gene Ontology analysis of the Pol II clusters was done using the online tool G:Profiles (Reimand et al., 2007). The Ensembl transcript IDs corresponding to the TSSs of the different clusters and expression states were used as the input. G:Profiler then produced a list of GO terms that were enriched or depleted in one combination of cluster and expression status compared to the others.

**Comparison with previous promoter classification**

A classification of *D. melanogaster* promoters into three classes based on the pattern of sequenced capped mRNA fragments around around the TSSs was downloaded from a published study (Ni et al., 2010). The match between the promoters following this classification and those
classified according to the Pol II clusters was done using their unique Ensembl transcript IDs.

A.5 METHODS FOR CHAPTER 5

A.5.1 Creation of Capture Hi-C interaction lists

Preparation of CHi-C libraries

Hi-C libraries were created from murine embryonic stem cell (mESC) and fetal liver (FL) samples by Stefan Schoenfelder and from human CD34+ and GM12879 cells by Cameron Osborne using the protocol from Lieberman-Aiden et al. (2009) with small modifications. The restriction enzyme used in all four conditions was HindIII. Duplicates were created for the mESC, FL and CD34+ samples, and triplicates were created for the GM12878 samples.

Enrichment of the Hi-C libraries to create promoter Capture Hi-C libraries was done using SureSelect target enrichment and a custom-designed bionylated RNA bait library from Agilent Technologies targeting HindIII fragments containing promoters as defined by the Ensembl gene start coordinates. The resulting libraries were then sequenced on the HiSeq1000 Illumina platform using paired-end sequencing.

Read processing and interaction calling

The processing of CHi-C reads was done by Bori Mifsud and Robert Sugar of the Luscombe Laboratory. The reads were mapped using the HiCUP pipeline, which mapped the reads to the reference human and mouse genomes, filtered circularised reads and re-ligations and removed all duplicated reads. An extra filtering step removed do-tags that did not contain at least one fragment targeted by the capture step.

Significantly interactions in each replicate were called using the GOTHiC Bioconductor package. This algorithm assumes that the biases in Hi-C experiments is captured in the mapped read coverage. Significantly interacting pairs of regions were identified using a cumulative binomial test followed by Benjamini-Hochberg multiple testing with a 0.05 FDR cutoff. Interactions between fragments less than 20 kb apart were discarded, as were those whose interacting fragment’s coverage was higher than 3x and less than 0.2x of the median coverage. Finally, only those interactions considered significant at least two replicates were considered for downstream analysis.

The manipulation of genomic interactions was done using a GenomicInteractions R class created by Robert Sugar from the Luscombe Laboratory, based on the Bioconductor GenomicRanges package (Aboyoun et al., 2013).
A.5.2 Expression analysis

The classification of baited fragments into expression classes was performed by Bori Mifsud from the Luscombe Laboratory. Genes were separated into 5 expression categories based on mRNA-seq data. The datasets used were those with the Gene Expression Omnibus IDs GSM723776 (Shen et al., 2012) and ERR031629 (Ficz et al., 2011) for ESC cells, GSM661638 data from Ter119+ cells for FL (Kowalczyk et al., 2012), GSM758572 and GSM758559 ENCODE data for GM12878, and GSM981257 ENCODE data for CD34+ cells.

RPKM values were obtained from the mRNA-seq datasets using TopHat with default settings (Trapnell et al., 2009). Genes with 0 RPKM formed a separate category and all other genes were divided into quartiles according to their RPKM values. Should a bait fragment overlap with more than one Ensembl promoter and should those promoters be assigned to a NA expression class. Otherwise, should all promoters overlapping with a bait fragment belong to the same expression class or should only one promoter overlap with the bait fragment, the bait fragment is assigned to that expression category.

A.5.3 Interactions with enhancer-annotated fragments

HindIII fragments were classified as containing an enhancer if they showed any overlap with sets of previously published enhancers for the mESC and FL cells (Shen et al., 2012), with the enhancer classes from the ENCODE genome segmentation analysis for the GM12879 cells (ENCODE Project Consortium et al., 2012), and with genomic regions provided by Bori Mifsud with a high signal for the histone modifications H3K4me1 and H3K27ac according to the genome segmentation tool ChromHMM (Ernst et al., 2011) for the CD34+ cells.

After identifying the enhancer-containing fragments, the number of such fragments that each bait fragment contacts was tabulated and compared to the expression class of the bait fragments. The results were visualised as bar charts representing the proportion of bait fragments contacting 0, 1, 2 to 9 or more than 9 enhancer-containing fragments that belong to each expression class.

A.5.4 Genomic interactions and chromatin modifications

Histone modification datasets used

Several datasets detailing the genomic coordinates of histones modifications were downloaded from publicly available sources. The detailed lists of datasets used, their Gene Expression Omnibus IDs and their publications are presented in tables 17, 18, 19 and 20.
Identification of the genomic localisation of histone modifications

The datasets used for comparison in the GM12878 cell line were produced as part of the ENCODE Project. For these datasets the processed data containing the genomic intervals that were enriched for the histone modifications was downloaded from the consortium website. For the other three cell types and for those datasets where raw reads were available, these were mapped to the reference mouse or human genome using Bowtie \(^{[\text{Langmead et al., 2009}]}\), allowing for one mismatch in a seed of 25 and filtering reads that map to more than one location on the genome. The command line used was as follows:

```
bowtie -n 1 -l 25 -m 1 <genome_reference_file> \
    <sequenced_reads.fastq> <output_file.bowtie>
```

The mapped IP and input reads were then used to call peaks using MACS \(^{[\text{Zhang et al., 2008}]}\) with all default settings. The final output was a series of genomic intervals defining the areas enriched for the histone modifications.

**Enrichments of histone modifications in non-bait fragments**

HindIII fragments were defined as containing a certain histone modification if there was any overlap between the genomic intervals of the restriction enzyme fragments and those enriched for a histone modification. The enrichments for each chromatin modification in the sets of non-bait fragments interacting with the baits from each expression class were calculated using the following equation:

\[
\text{enrichment} = \log_2 \left( \frac{\frac{\text{FragmentsInExpressionClassOverlappingMark}}{\text{FragmentsInExpressionClass}}}{\frac{\text{Non-baitFragmentsOverlappingMark}}{\text{Non-baitFragments}}} \right)
\]  

(4)

The resulting enrichment values are positive if there is an enrichment and negative if there is a depletion of the histone modification. In order to unequivocally assign a non-bait interacting fragment to an expression class, only those fragments that interact with bait fragments in the same expression class were used. The visualisation of the enrichments was done using separate heat maps for each cell type. In addition to the enrichments on all interacting non-bait fragments, separate enrichments were calculated for those interacting non-bait fragments that overlapped with an annotated enhancer and for those that did not.

**Distance controls in histone enrichments**

In addition to the histone enrichments heat maps that were created using the full interaction datasets, other heat maps were created using only the subsets of interactions that spanned over set distance thresholds. The distance spanned in an interactions was measured as the
difference in genomic coordinates between the midpoint of the bait fragment and the midpoint of the interacting non-bait fragment. Interchromosomal interactions were considered to span a positive infinite distance.

The thresholds used for the distance controls were 100 kb, 500 kb and 1 Mb. Apart from subsetting the list of interactions to those that spanned over the selected thresholds, the enrichment calculations were done in a manner identical to those of the full dataset. The distance controls for each sample were plotted together with the original full dataset enrichments so that the same scale allows for easy comparison of the results.

Enrichment of chromatin states in non-bait fragments

The genome annotation of chromatin states in the human GM12878 cell type was downloaded from ENCODE (ENCODE Project Consortium et al., 2012), in the form of the genomic intervals that are classified as belonging to one of the chromatin states. A set of chromatin states was created for the mESC cell type by running ChromHMM (Ernst and Kellis, 2012) using default parameters on datasets from ChIP-seq experiments done in this cell type. A range of possible number of states was used, and the characterisation of seven states was used as the one that captured the most biologically relevant combinations of TFs and histone modifications.

The enrichment calculations were done in the same manner as the enrichments of histone modifications, using equation 4. A HindIII fragment was considered to belong to a state if any of it overlapped in any coordinate with the one of the chromatin state’s genomic intervals.

Cohabiting marks in interacting fragments

The association of any two chromatin marks in non-bait fragments that interact with bait fragments was measured as the proportion of fragments that show any overlap with histone mark X that also show an overlap with another mark Y. The resulting values range from 0 if there are no fragments that contain both marks, to 1 if all fragments that contain mark X also contain mark Y.

Enrichments of histone modifications in promoter-promoter interactions

The same set of chromatin modifications used for the analysis of enrichments in non-bait interacting fragments was used for the enrichments among promoter-promoter interactions. The determination of whether a bait fragment contains a histone modification was done in the same way as for non-bait fragments: if the bait fragment overlaps any of the coordinates of the histone modification it was considered as containing that histone modification.
The enrichments were obtained by measuring how likely a modification $i$ was to be in a promoter interacting with another promoter that contained the modification $l$ compared with how likely the modification $i$ was to be seen in any of the promoter-promoter interactions. The results for this analysis in the mESC and GM12878 cell types were plotted in heat maps.

A.5.5 Interactions in TADs

As a measure of directionality, for every bait-fragment that fell within an annotated TAD the extent to which interactions involving that promoter connected with a fragment that fell further along the chromosome instead of before it was calculated. The value obtained was the proportion of fragments that interacted with a bait fragment that had a higher genomic coordinate than that bait fragment, divided by the total number of fragments that interacted with it. This proportion was then scaled to a -1 to 1 interval.

The genomic coordinates of topologically associated domains (TADs) were downloaded from a previous study that performed Hi-C in human and mouse cells (Dixon et al., 2012). The relative positions of all bait fragments compared to the TAD coordinates were calculated. Finally, histograms were plotted showing the directionality of the interactions from bait fragments that fell within each tenth of the TADs. This analysis was performed separately for all interactions, for those between promoter and non-promoter fragments and for those involving two promoter fragments.

A.5.6 CTCF and cohesin analysis

Preprocessing of datasets

The GM12878 CTCF and cohesin datasets used in this analysis were downloaded as peak coordinates from the ENCODE Project consortium website. For the mESC analysis, raw reads were downloaded and the peaks were identified in the same manner as the histone modifications used in the C-Hi-C analysis.

CTCF bridging by interactions

A CTCF was considered bridged by an interaction if a baited fragment that did not overlap with the CTCF peak contacted a non-baited fragment whose genomic coordinates were past those of the CTCF peak. In the same manner, an interaction was considered to bridge a CTCF peak if there was a CTCF peak in the interval between the non-bait fragment and the bait fragment it interacted with. Using these definitions the proportion of CTCF peaks that were bridged by an interaction was calculated for both the GM12878 and mESC cell types.
A.5.7 Interactions and LADs

The set of lamin-associated domains (LADs) used in this analysis was taken from a previous study in mouse cells (Peric-Hupkes et al., 2010). Every interaction between a bait and a non-bait fragment was classified according to the location of the bait and non-bait fragment relative to the LADs. If any part of a fragment overlapped with any of the LAD coordinates, that fragment was considered as belonging to the LAD. This way, interactions were separated as not being involved in LADs, as joining two fragments inside the same LAD (Intra LAD), as joining two fragments in different LADs (Inter LAD), as having the bait fragment in a LAD and the non-bait outside it (Bait in LAD) or vice-versa (Non-bait in LAD). The number of interactions in each category was then plotted according to their expression class.