A computational study of transcriptional regulation in eukaryotes on a genomic scale

Florence Marie Géraldine Cavalli

Darwin College

University of Cambridge

This dissertation is submitted for the degree of Doctor of Philosophy

December 2010

Preface

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.

Statement of length

This thesis does not exceed the length limit specified by the Degree Committee for Biology, which is 300, single-sided, pages of double spaced text, not including the bibliography and appendices.

Summary

Precise regulation of gene expression is essential in higher eukaryotes: only by producing the correct cellular components can cells perform their intended function correctly. Transcription is a complex process that is regulated by a combination of factors including DNA sequence and its three-dimensional topology, transcription factor proteins (TFs), co-factors and histone modifiers. The computational work in this thesis focuses on trans-acting factors of eukaryotic regulation: (i) at the level of TFs by analysing their expression across different human tissues and identifying sets of factors that potentially confer tissue-specific expression, and (ii) at the level of histone modification by examining the effects of the acetyltransferase protein males-absent-on-the-first (MOF) in *Drosophila melanogaster*.

The thesis consists of six chapters. In Chapter 1, I introduce transcription and its regulation in eukaryotes, discussing some of the methods employed to study these processes. In Chapter 2, I present a new statistical method for processing microarray data that detects which genes are specifically expressed in a small number of conditions. I apply the method to a dataset of 32 human tissue samples and demonstrate its utility by identifying both previously known and novel tissue-specific genes. In Chapter 3, I use this analysis to examine the expression of human TFs and identify those that are potentially responsible for controlling tissue specificity. I determine the TFs that are ubiquitously expressed across all tissues and also highlight the ones that are expressed specifically in particular tissues. Using co-expression and protein-protein interaction data, I then build a network of TFs that might function in a combinatorial manner. Using this network, I examine the relationships between tissue-specific and ubiquitous TFs and relate groups of TFs to particular biological functions.

In Chapter 4, I describe an investigation into the global effect of histone acetylation on transcription done in collaboration with Dr Asifa Akhtar's group at the

Max-Planck Institute of Immunology in Freiburg (Germany). We use dosage compensation in *D. melanogaster* as a model for chromosome-wide transcriptional regulation. This system up-regulates gene expression on the male X chromosome in order to correct for the imbalance in gene dosage resulting from the absence of one X chromosome in male cells relative to autosomes and female cells. The process is controlled by the histone-modifying enzyme MOF which acetylates lysine 16 of histone H4 (H4K16), and causes the up-regulation of the entire male X-chromosome. Using data from chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-Seq) experiments, I identify the binding patterns of MOF on the fly genome and assess the amount of H4K16 acetylation in the neighbourhood of the MOF-binding sites. I then determine the effect of MOF-binding and acetylation on gene expression by examining transcriptomic data from mutant flies lacking a functional MOF gene. In Chapter 5, I extend this analysis by measuring the influence of MOF on RNA polymerase II activity. Comparison of ChIP-Seq data from male and female flies reveals that RNA polymerase II-binding is increased in the body of dosage-compensated genes in males, showing that MOF enhances the production of mRNAs by increasing the amount of RNA polymerase II going through dosage-compensated genes. Finally in Chapter 6, I discuss the importance of my findings for understanding transcriptional regulation and highlight gaps that remain in our knowledge.

In summary, this thesis identifies TFs that are responsible for tissue-specificity in humans and provides candidate sets of TFs involved in combinatorial regulation. Additionally, it shows the effect of a histone-modifying enzyme on polymerase activity on a genome-wide scale and contributes to the understanding of the role of histone modifications in global transcriptional control.

Acknowledgements

Firstly, I would like to thank my supervisor Nicholas Luscombe for his encouragement, support and creating a favourable research environment throughout my time at the EBI.

Working on this thesis has been a great journey and I have enjoyed the company and support of many people. My gratitude goes out to these people who contributed to this learning experience.

My thanks go to the past and current members of the Luscombe group. Having been great friends and colleagues during these four years. They have also helped created a good research environment and given me essential input and feedback on my work. So my thanks go to Kathi, Filipe, Judith, Aswin, Inigo, Anabel, Juri, Corina, Karthi, Garth and Bori. I specially thanks Juanma for his tremendous support and guidance during this four years. This thesis would have not been possible without him.

I would like to thank Richard Bourgon who helped me with kindness and wonderful support in statistics. Thomas Conrad and Asifa Akhtar from MPI-Freiburg for offering me the opportunity to build a very productive collaboration.

Further, I would like to thank the "Predocs" and other "Postdoc" friends, for being so friendly and contributing immensely to make me enjoy my life during this four years in Cambridge.

I also would like to thank other friends for their true friendship and support all along my PhD; Catherine & Nicolas, Isabelle and Claire.

I am very grateful for the new support and love from Lars.

Finally, I am especially thankful to my brother and my parents for their love and unfailing support without them I would have not achieved what I have until now.

Cambridge, December 2010

Contents

List of Figures

List of Tables

Abbreviations

Ac:	Acetylation
bp:	Base Pairs
BIC:	Bayesian Information Criterium
C. elegans:	*Caernorhabditis elegans*
CES:	Chromatin Entry Sites
D. melanogaster:	*Drosophila melanogaster*
DC:	Dosage Compensation
DCC:	Dosage Compensation Complex
EM:	Expectation-Maximisation
FDR:	False Discovery Rate
HAT:	Histone Acetyl Transferase
HP1:	Heterochromatin Protein 1
H4K16ac:	Histone H4 Acetylated on lysine K16
Log2FC:	Log2 Fold-Change
M2Y:	Mammalian-2-Hybrid

Me:	Methylation
MLE:	Maleless
MOF:	Males-absent-On-the-First
MSL:	Male-Specific Lethal complex
MSL1:	Male-Specific Lethal 1
MSL2:	Male-Specific Lethal 2
MSL3:	Male-Specific Lethal 3
NSL:	Non-Specific Lethal
Ph:	Phosphorylation
PIC :	PreInitiation Complex
Pol II:	RNA polymerase II
PPI:	Protein-Protein Interaction
PWM:	Position Weight Matrix
ROC:	Receiver Operating Characteristic
StI:	Stalling Index
TF:	Transcription Factor
TFBS	Transcription Factor Binding Site
TiGER:	Tissue-specific Gene Expression and Regulation
TSS:	Transcription Start Site
wt:	Wild Type

1

Introduction

All cells in a multicellular organism contain the same genetic information; however each organism consists of a large array of cell types that perform diverse biological functions. This cell type diversity results from differences in gene expression, the process by which information encoded in DNA is transcribed to RNA, and then in many cases, translated to proteins that are used by the cell to perform specific functions. Importantly, gene expression is tightly regulated to produce the right amount of protein at the right time in each cell. The focus of this thesis is the genome-scale analysis of eukaryotic transcription and its regulation, using computational methods.

This first chapter describes the general principles of gene expression and discusses the importance of its precise regulation. It then outlines the mechanisms for eukaryotic transcription and its regulation. The experimental datasets, computational databases and methods used in this work are introduced. The chapter concludes with a summary of the results presented in the thesis.

1.1 Gene expression

1.1.1 From DNA to proteins, a well-regulated process

Many processes take place in order to produce the final protein produced from the genomic DNA in eukaryotic cells. These processes are highly regulated at different steps, and allow cells to control the set of RNA and proteins that are present in a dynamic manner. First, transcription produces RNAs by the action of the RNA polymerases that read the genetic information encoded in DNA. This involves the precise recruitment of many proteins and enzymes in addition to the polymerase. Polymerase function is regulated through local and specific alteration of chromatin structure - such as nucleosome displacement or histone modifications - as well as the cooperative action of many general and specific regulatory proteins. Next, splicing, which is often coupled with transcription, removes non-coding introns from pre-mRNAs. Different exons are selectively included in the final transcript depending on the action of sequence-specific RNA-binding proteins. The abundance of transcripts is affected by the rate of degradation, as well as production, and further processes such as mRNA-capping and polyadenylation enhance the stability of new transcripts. Many non-coding RNAs will remain in the nucleus and function as regulatory RNAs that affect transcription itself. Mature mRNAs on the other hand, are exported outside the nucleus to be translated by ribosomes in the cytoplasm or the endoplasmic reticulum. Finally, the resulting polypeptide chains are folded and often undergo post-translational modifications to become functional.

Among all the processes described above, transcription represents the first step for the production of functional proteins in the cell. Its control is arguably the most important point of control, and a properly functioning transcriptional regulatory system is essential for cell viability.

1.1.2 Mis-regulation of gene expression in disease

Changes in gene expression have been associated with many diseases. Studies comparing diseased and normal cells have revealed that gene expression levels are often perturbed. For example, cancer cells have been shown to express different genes compared with normal cells, which both arises as a result of irregular cellular functions and also in turn contribute to further abnormalities. Abnormal transcript level can appear for several reasons: for example, it could result from the mis-regulation of transcription, or expression levels might be altered owing to the presence of extra copies of particular genes. Indeed, karyotypes of some cancers reveal extra copies of entire or part of chromosome(s).

Further, mutations of transcription factor-coding genes and alterations of the regulatory target sequences have been identified as contributors to developmental disorders (reviewed in Maston et al., 2006; Engelkamp and van Heyningen V., 1996). Transcription factors are notably overrepresented among oncogenes (Darnell, 2002; López-Bigas et al., 2006; Furney et al., 2006); for instance, ATBF1 is implicated in human prostate cancer (Sun et al., 2005b).

Therefore it is important to increase our understanding of how gene expression is regulated, as this is likely to reveal new targets for treating diseases.

1.2 The eukaryotic transcription

1.2.1 General view

Transcriptional regulatory mechanisms modulate the accurate recruitment and activation of RNA polymerase at different locations in the genome, most typically gene promoters. This is mediated at different levels (Figure 1.1), which could be viewed as distinct but interacting modules of regulatory organisation: for example in

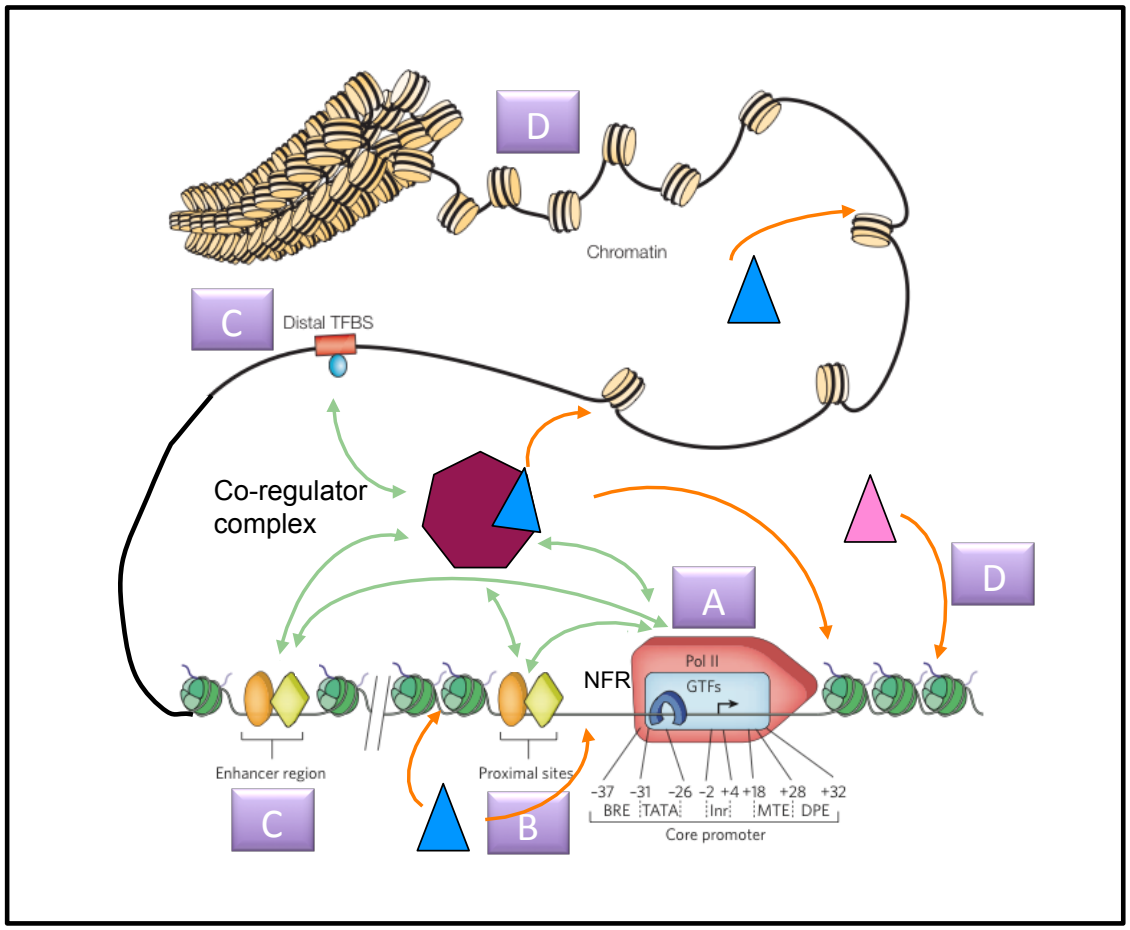


Figure 1.1: Transcriptional regulation by modulation of RNA polymerase II function. Different levels at which transcription is regulated: (A) promoter level: the general transcription machinery (composed of Pol II and general TFs) is assembled and readied for activation at the promoter by the action of general transcription regulators, (B) promoter proximal level: sequence-specific DNA-binding TFs (lozenge and oval) bind at proximal sites working directly or with co-regulator(s) to act on the general transcription machinery, (C) enhancer level: sequence-specific DNA-binding TFs bind at enhancer region(s) acting directly or in collaboration with other TFs or co-regulators on the general transcription machinery, (D) nucleosome level: chromatin modifiers that act on nucleosomes such as nucleosome remodelers or histone modifiers (triangles). (Adapted from Wasserman and Sandelin, 2004; Fuda et al., 2009).

cis, different regulatory events take place at promoter-proximal transcription factor binding sites, at enhancers and at entire groups of nucleosomes to contribute to the successful transcription of a single gene. In trans, protein complexes interact with regulatory regions and play a critical role in transcriptional regulation.

Transcriptional regulators can be grouped into three types. First, the preinitiation complex (PIC) binds at the core promoter and recruit Pol II (Figure 1.1 – A). Second, DNA-binding transcription factors bind to sites such as proximal pro-

moter elements and enhancers (Figure 1.1 –B & C). Third, enzymes modify the higher-order chromatin structure by promoting the physical movement of nucleosomes relative to the genome and post-translationally modifying histone molecules to alter the stability and accessibility of chromatin (Figure 1.1 -D). Thus, in the sections below, we describe the general principles of Pol II function. We then discuss transcription factors (TFs) and the contribution of histone modifications to transcriptional control.

1.2.2 RNA polymerases

RNA polymerases are large multi-subunit enzymes that perform transcription and thus produce all the RNAs in the cell from a DNA template. These complexes are assembled and tightly bound at the promoter of genes before initiating transcription. There are three types of RNA polymerases in eukaryotes, each synthesizing different classes of RNAs: (i) RNA polymerase I which synthesises most ribosomal RNA, (ii) RNA polymerase II which synthesises all protein-coding mRNAs, miRNAs and some small RNAs (e.g, those in spliceosomes), (iii) RNA polymerase III which synthesises tRNAs, the 5S rRNA and many other small RNAs.

The RNA polymerase structure is broadly conserved from prokaryotes to all eukaryotes (Saltzman and Weinmann, 1989). All three RNA polymerase types are believed to have diverged from a common ancestral protein, as some subunits are shared between them. The largest subunits of the three eukaryotic RNA polymerases show significant similarity. Here, we will focus on RNA polymerase II (Pol II) and its regulation, as it is at the origin of the production of all proteins in the cell.

1.3 The eukaryotic transcriptional process and its regulation

1.3.1 RNA polymerase II transcriptional mechanism

1.3.1.1 RNA polymerase II transcriptional steps

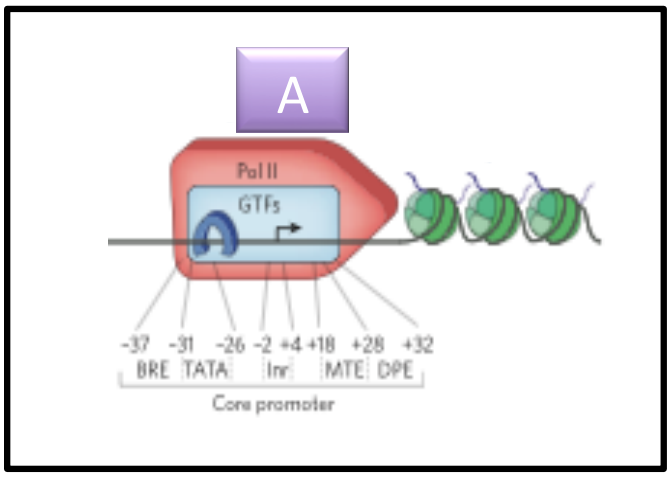


Figure 1.2: Transcriptional regulation by modulation of RNA polymerase II function -A. RNA polymerase II and general TFs (GTFs) binding at the core promoter forming the general transcription machinery. (Adapted from Fuda et al., 2009).

Pol II must assemble at the promoter region of genes in order to perform transcription. The core promoter is a small section of DNA encompassing the transcription start site (TSS) that serves as the docking site for the basic transcriptional machinery and PIC. The main elements that have been identified on this site are: the TATA box, Initiator Element, Downstream Core Element, TFIIB-Recognition Element, Motif Ten Element and Downstream Promoter Element. With the exception of the TFIIB-Recognition Element which is bound by TFIIB, all the other elements have been described to interact with TFIID (reviewed in Smale and Kadonaga, 2003; Maston et al., 2006). Single promoters do not contain every single element listed here, but different promoters can encode for distinct combinations of these

elements. Further many promoters do not contain any of the consensus elements that have been described so far.

Pol II assembly at the core promoter requires numerous general transcription factors. These proteins form the basis of the pre-initiation complex and are believed to be responsible for the positioning of Pol II at the core promoter region. They have been identified by isolation, co-purification and characterised by mutagenesis or through assays measuring Pol II activities (Hori and Carey, 1994).

Pol II activation requires DNA-binding of general transcription factors (such as TFIID, TFIIB, TFIIE, TFIIF, TFIIH) that form the PIC. Transcriptional activators and co-activators then interact with the complex to initiate transcription. Following transcriptional activation, the elongation takes place with the aid of elongating factors. Finally, the process terminates after transcription of the poly(A) site. This is believed to occur as a result of a conformational change in Pol II after transcribing the poly(A) site which makes the enzyme susceptible to termination, and/or due to the action of an exonuclease that degrades the RNA still attached to Pol II after cleavage and release of the transcript. In summary, transcription takes place through the recruitment and action of multiple regulatory factors.

In parallel to the action of general transcription factors and cofactors on Pol II, the enzyme undergoes several post-translational modifications. Since these modifications dictate the functional state of the polymerase, distinct transcriptional phases can be identified by the amino acid residues that are phosphorylated (reviewed in Hirose and Ohkuma, 2007). These modifications occur on the carboxy-terminal domain (CTD) tail of the largest Pol II subunit Rpb1. First, Pol II is recruited to the gene promoter in a hypo-phosphorylated state for assembly. Second, its release from the promoter occurs when TFIIH (CDK7 in *Drosphila*) phosphorylates Ser5 (Ser5P) of the CTD. Third, Pol II activity can be paused through the action of negative factors. Finally, elongation begins when the transcription elongation factor b

(p-TEFb; CDK9-cyclin T in *Drosphila*) phosphorylates Ser2 (Ser2P) (Peterlin and Price, 2006) (Figure 1.3). This last modification is also linked to co-transcriptional RNA-processing as it helps recruit splicing and polyadenylating factors. At termination, the CTD modifications are then reverted by phosphatases, making the Pol II ready for another round of transcription. Some modifications, such as Ser5P and Ser2P, are well characterised, while others, such as Ser7P, are still not completely understood. There is potential for combinatorial activity and an increased knowledge of how these modifications alter the behaviour of Pol II will greatly improve our understanding of transcription at a mechanistic level.

The fact that these Pol II modifications are closely linked to different transcriptional states, and the use of antibodies that are specific for particular modifications allow us to identify Pol II at different stages of transcription. Chromatin immunoprecipitation (ChIP) data measuring Pol II-binding on the *Drosophila melanogaster* (*D. melanogaster*) genome are analysed in chapter 5, during which we study the effects of histone acetylation on Pol II function.

1.3.1.2 RNA polymerase II transcription dynamics

After initiation, transcription generally continues into elongation; however, if the transcription stops after only 20 to 50 nucleotides Promoter Proximal Pausing (PPP) happens. Here, Pol II accumulates at the 5'end of genes while still retaining their elongation potential (Core et al., 2008). It was first observed at the promoter of the heat-shock gene Hsp70 in *D. melanogaster* (Gilmour and Lis, 1986). At this promoter Pol II is recruited before activation, the pre-initiation complex is then prepared for rapid activation upon heat shock (Boehm et al., 2003). More generally, pausing can affect initiated Pol II to different extents: (i) by preventing the full-length transcription of genes on which Pol II has been initiated, or (ii) by regulating the release of Pol II into productive elongation of active genes (Core and Lis, 2008).

Performing ChIP-chip experiments in *D. melanogaster*, Zeitlinger et al. (2007) and Muse et al. (2007) demonstrated that polymerase pausing is a genome-scale phenomenon affecting $\approx$ 30% of genes. Similar observations were made in mammalian systems by Guenther et al. (2007) and Core et al. (2008). This last study used a new method called GRO-seq to detect transcriptionally engaged Pol II in primary human lung fibroblasts. Importantly, pausing was observed over a large dynamic range or transcript production, and both highly and lowly expressed genes are regulated at this level. Among the paused genes identified in *D. melanogaster* studies, there was a significant over-representation of genes involved in developmental regulation and cell signalling. This suggests that regulation through pausing is a fundamental step for controlling developmental programs and enabling rapid reaction to environmental stimuli (Zeitlinger et al., 2007; Wang et al., 2007).

Additionally, another role of pausing was recently hypothesised by Gilchrist *et al.*. Performing microarray analysis of Negative Elongation Factor (NELF)-depleted drosophila S2 cells (a negative regulator of elongation), they observed up-regulation of genes, but also unexpectedly, down-regulation. They hypothesised that Pol II pausing positively regulates expression by maintaining accessibility of the promoter (Gilchrist et al., 2008). In the last few years, more and more evidence has emerged showing that regulating Pol II escape from pausing is a universally important control point that prepares genes for future activation (Gilchrist et al., 2008; Lee et al., 2008; Price, 2008; Gilmour, 2009).

In summary, Pol II activity can be regulated at a basic level through assembly at the promoter, post-translational modification, and release into elongation. However, these steps only enable low levels of transcription, and further additional levels of regulation also modulate polymerase function.

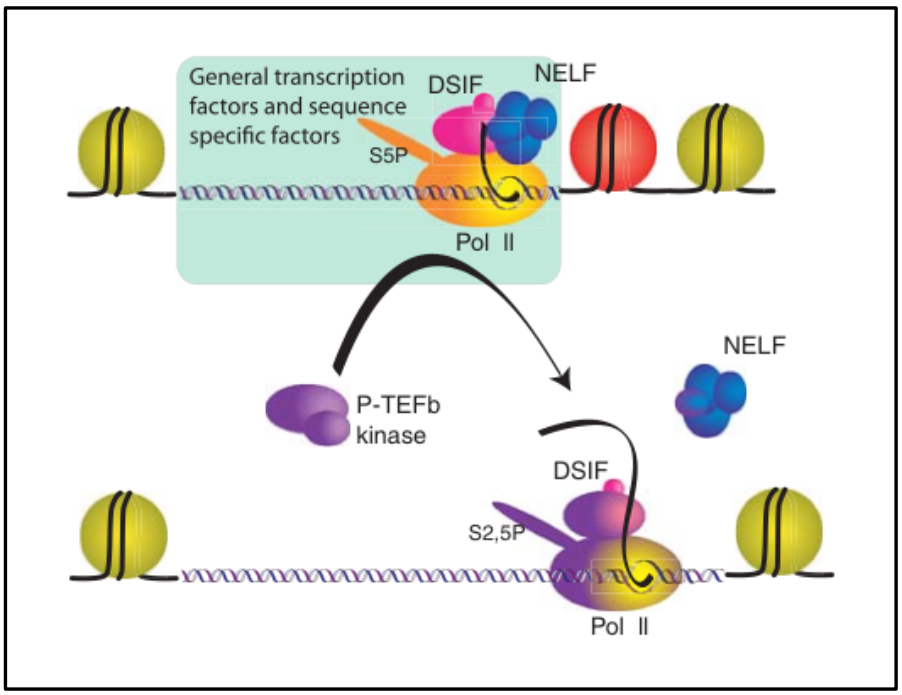


Figure 1.3: Pol II transcription dynamics, promoter proximal pausing. Pol II initiates transcription but pauses after 20 to 50 nucleotides. Serine 5 (S5P) of the CTD tail are phosphorylated by TFIIH (top panel). Negative elongation factor (DSIF and NELF) associate with the Pol II complex and cause Pol II to pause. The release of the paused Pol II is mediated by phosphorylation (designated by purplish hue) of NELF, DSIF, and serine 2 of the CTD by the kinase P-TEFb (bottom panel). Figure from Gilmour, 2009.

1.3.2 How transcription factors regulate transcription

1.3.2.1 The transcription factors

Transcription Factors (TFs) are proteins that regulate transcription by influencing the recruitment of Pol II at promoters. These proteins generally contain DNA-binding domains, but additionally many have an effector domain that influences Pol II activity. In general, TFs proteins bind to specific DNA sequences at promoters or enhancers. They contain one or more DNA-binding domains that are able to recognise short, specific DNA sequences. These DNA sequences are referred to as transcription factor binding sites (TFBS); they are typically 6-12 bp long and their sequence content are often represented using a position weight matrix (PWM). These matrices show the consensus DNA sequence in which certain po-

sitions are constrained and others are more variable depending on the effect they have on TF binding. The efficiency of TF binding can be affected by variation in the TFBS sequence, as it directly affects the strength of binding. On the TF-side, amino acid changes in the DNA-binding domain can alter its target-sequence. TFs largely bind in the major groove of the DNA forming a molecular interaction between amino acid side-chains and nucleotide base edges (Luscombe et al., 2000). TFs are usually grouped in families based on the structure of their DNA binding domains. This is biologically relevant because some binding domains are linked to the function of their targets. For example homeodomains are generally linked to developmental processes (Driever and Nüsslein-Volhard, 1989) and factors from the interferon regulatory factor family trigger immune response against viral infection (Luscombe et al., 2000). However, other families such as C_2H_2 zinc-finger protein family are involved in numerous cellular processes (Wu, 2002). The modular nature of the C_2H_2 DNA binding domains gives them more flexibility in DNA recognition and interactions with other proteins, which in turn offer them the possibility to regulate a wider range of processes.

Identifying the repertoire of TFs encoded in a genome is an important step towards understanding the organism's regulatory system. Recently, a high-quality dataset of 1,391 DNA-binding TFs in the human genome was published by Vaquerizas et al. (2009). Only few of these TFs are functionally characterised so far. Some DNA-binding domains are more prevalent than others. Indeed, the C_2H_2 zinc-finger, homeodomain and helix-loop-helix families represent 80% of the human TF repertoire. The C_2H_2 zinc-finger proteins are stabilized by a zinc ion that is coordinated by two cysteines and two histidines. Homeodomains have a helix-turn-helix DNA-binding region and members of this family are essential for formation of the anterior-posterior body axes throughout the animal kingdom. Helix-loop-helix proteins contain a basic region for DNA binding and a neighbouring region for

dimer-formation.

Many TFs function as homo- or hetero-dimers formed with other members of the same family (for example, helix-loop-helix proteins). Heterodimerisation allows recognition of a large number of DNA-sequence targets using a relatively small set of family members. It also allows bipartite binding sites (two binding-sites separated by space) to be recognised (for example the Fos-Jun heterodimer, Risse et al., 1989). Finally, TFs can act as activators, repressors or both depending on the context in which they function (for example, the presence of other regulatory proteins).

1.3.3 Action of transcription factors

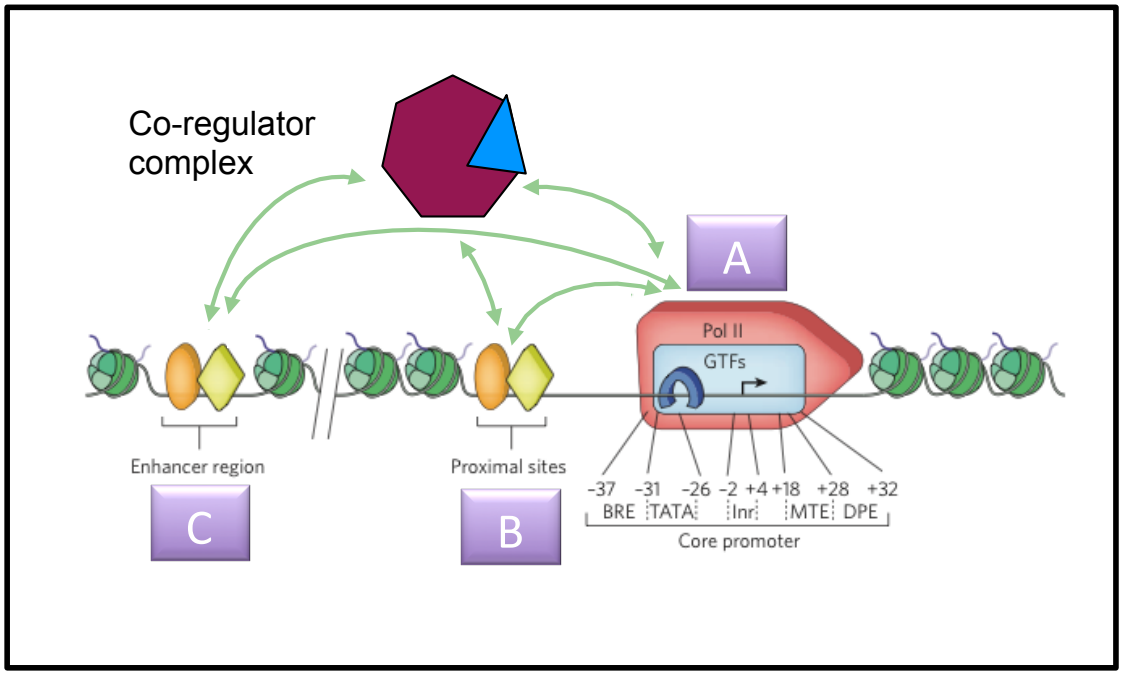


Figure 1.4: Transcriptional regulation by modulation of RNA polymerase II function - B & C. Action of TFs on RNA polymerase II. General TFs (A) are part of the general transcription machinery. Sequence-specific DNA-binding TFs bind to proximal promoters (B) or enhancers (C). TFs act directly or indirectly—via co-regulator(s) such as the mediator complex—on the RNA polymerase II to regulate the transcription. (Adapted from Fuda et al., 2009).

TFs are divided into two main classes: general and sequence-specific DNA-binding TFs. General TFs are components of the pre-initiation complex and they are necessary for basal RNA polymerase activity (see section 1.3.1.1). Sequence-

specific DNA-binding TFs interact with the core transcription machinery and are needed for condition-specific regulation of gene expression. They bind to specific binding sites in promoters or enhancers. Genes in higher eukaryotic genomes are likely to have several enhancers placed upstream and downstream, as well as within the introns and such regulatory elements can stimulate target transcription from a distance. Diverse mechanisms ensure that the right enhancer interacts with the right promoter; for example, an insulator protein can stop the action of a distal enhancer on a promoter if placed between them, or genes regulated by the same enhancers can compete for interactions with them through the use of tethering elements.

Combinatorial interactions between TFs and co-factors are important for correct gene expression. A repertoire of 1400 TFs is clearly not sufficient to control a complex gene expression program in humans, if the regulators act independently; however through combinatorial activity, they allow tight and precise spatio-temporal control.

A good example of combinatorial regulation can be seen in *D. melanogaster*, in which cis-regulatory modules (CRMs) integrate inputs from multiple TFs. By grouping several binding sites into each enhancer, they produce precise gene expression patterns with relatively few TFs. CRMs contain a mix of high- and low-affinity binding sites, this lead to their activation at different TF concentrations. Interestingly, it has been shown that in some cases activators are broadly expressed and activate their target genes. These genes can then be repressed only in well-defined region(s), where a repressor —binding on the same CRMs— is expressed. For example some stripes in drosophila embryos come from the expression, at a given place and time, of a repressor such as Slp1 (reviewed in Bonn and Furlong, 2008).

An interesting use of combinatorial regulation is the concept of a master regulator that can control the activity of a large number of CRMs at one time. If TFs are already bound to each CRM, a single master TF can have a global effect by

completing the combination across many CRMs. An example of a master regulator is MyoD, which controls muscle cell differentiation (Weintraub et al., 1989).

TF binding sites within enhancers are sometimes present in precisely spaced clusters, which allow TFs to bind in a specific manner. Stable assemblies of TFs are sometimes known as enhanceosomes, which can include not only transcriptional activators, but also structural proteins that stabilise the assembly of proteins (Merika and Thanos, 2001). Enhanceosomes are stable and expose a unique activating surface. They offer remarkable precision and efficiency for regulation by inducing the efficient loading of the basal transcriptional machinery. The virus-inducible enhancer of interferon-β is one of the best-understood enhanceosomes (Merika and Thanos, 2001). The combinatorial interaction between NF-κB proteins, the ATF-2/c-Jun heterodimer and the I(Y) group protein leads to a highly specific gene expression program during viral infections (Wathelet et al., 1998). The crystal structure of this enhanceosome was determined by Panne et al. (Panne et al., 2007).

To summarise, increasing knowledge has been obtained about how TFs function together (reviewed in Bonn and Furlong, 2008). However, there are still many TFs to be characterised and a lot to be discovered about their combinatorial action to regulate particular functions and cell types especially in higher organisms. Chapter 3 presents an analysis of human TF combinations in the context of tissue specificity.

1.3.4 How chromatin modifications regulate transcription

1.3.4.1 The structure of chromatin

The DNA molecule is not naked in the cell but bound by proteins resulting in a tight and efficient packaging. The combination of compacted DNA along with the protein components is called the chromatin. Chromatin can be restructured depending on environmental cues, to allow more or less access for regulatory proteins. Therefore

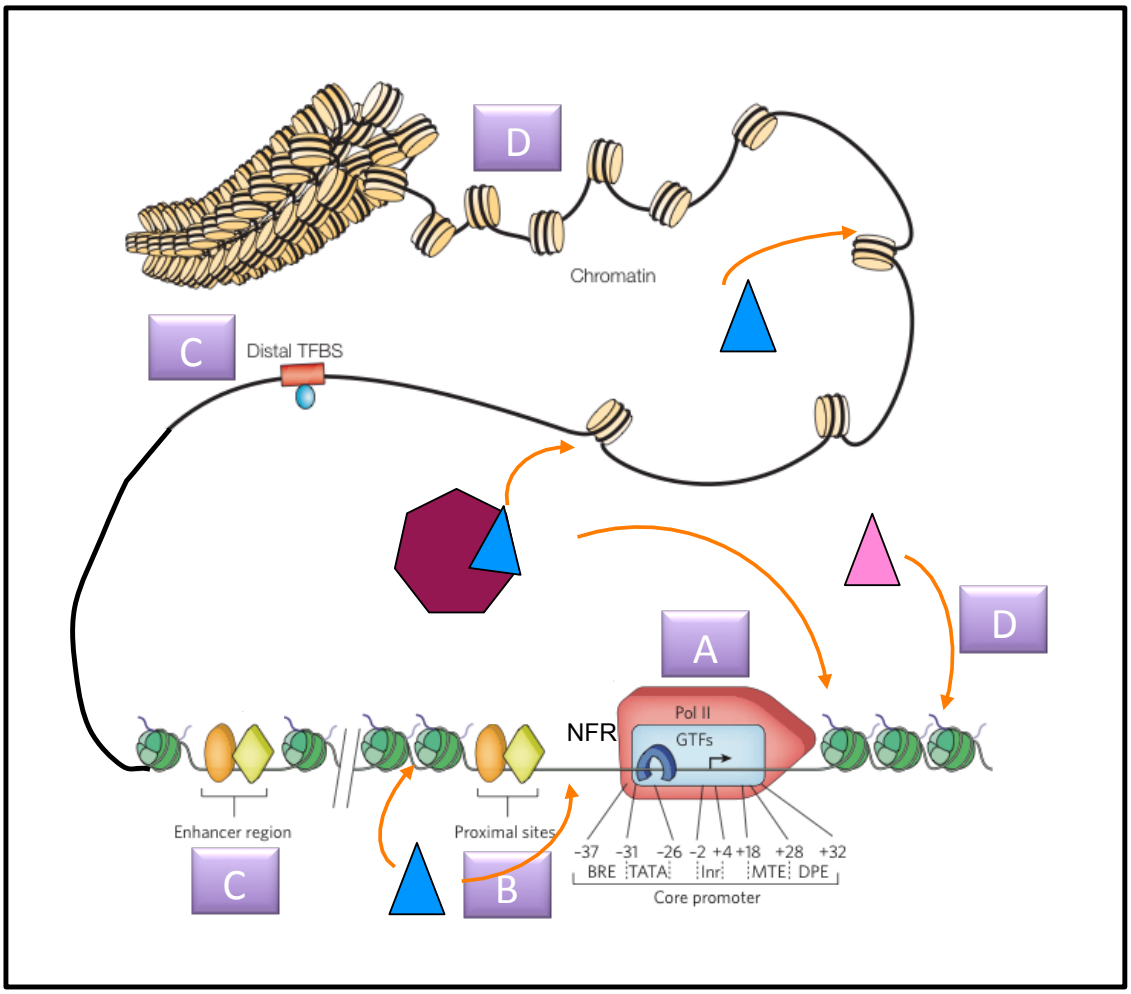


Figure 1.5: Transcriptional regulation by modulation of RNA polymerase II function - D. Chromatin modification influencing RNA polymerase II activity: Histone remodelling complex and histone modifying enzymes (triangles) act on the nucleosome to allow transcription by example creating the nucleosome-free region (NFR) or modifying the histone tails of nuclesomes present in the transcribed region (pink triangle). (Adapted from Wasserman and Sandelin, 2004; Fuda et al., 2009).

chromatin structure makes an important contribution to transcriptional regulation.

The basic unit of chromatin is the nucleosome: it consists of a 147bp segment of DNA wrapped around an octamer of histone molecules (H2A, H2B, H3, H4—two molecules of each histone). These core histone proteins are small and highly basic. Their amino acid sequences are highly conserved in all eukaryotes. Structurally, each protein contains a globular domain and a flexible "histone tail" protruding from the surface of the nucleosome (Figure 1.6). Chromatin at this level of compaction is also described as the 11nm fibre. Since nucleosomes do not provide enough compaction,

they are packaged further into a higher order structure called the 30nm fibre. This involves the linker histone H1 protein which is present in most higher eukaryotes —yeasts have a linker histone HHO1, which is thought to be orthologous to metazoan H1. Linker histone H1 binds to the segment of DNA linking two adjacent nucleosomes. At a higher order level of structure, we observe the 300-700nm fibre in the interphase and metaphase cell. The final chromatin structure can compact the original DNA by 10,000 fold of its original length.

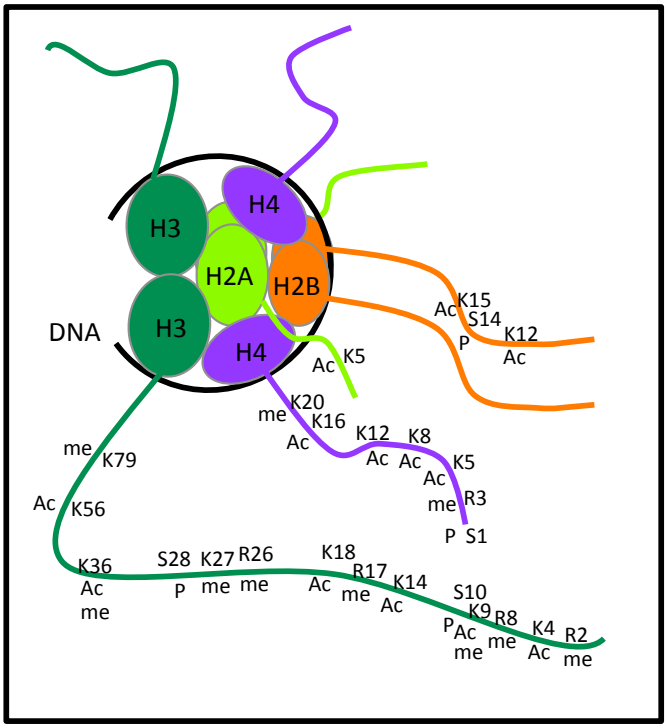


Figure 1.6: The nucleosome and associated histone modifications. Schematic representation of a nucleosome, octamer of the core histone proteins (H2A, H2B, H3, H4) around which DNA is wrapped. Histone molecules have long unstructured tails that can be covalently modified. All the known modifications for the Arginine (R) or Lysine (K) methylation (me), Lysine acetylation (Ac) and Serine (S) phosphorylation (P) are indicated. (Figure based on a figure from Allis et al., 2007)

The chromatin is generally characterised by two states; the euchromatin and the heterochromatin. Euchromatin or "active" chromatin is decondensed chromatin, consisting largely of coding sequences with the potential for transcriptional activity. This chromatin state undergoes many modifications through the action of many dif-

ferent proteins. For example, chromatin remodelling proteins utilise ATP to move a nucleosome along the DNA. In other cases, histone-modifying enzymes can introduce covalent modifications to specific histone residues. Heterochromatin can be defined as highly compact and silenced chromatin. It includes among other regions the centromeric and telomeric chromosomal domains and covers 96% of the mammalian genome. This state contains some well-known histone variants and histone modifications (see section 1.3.4.2). It is thought to have an essential role in the overall genome structure, expression and faithful chromosome segregation.

A recent study performed in *Drosophila* by Filion et al. (2010) refined this rather simple description of the chromatin states by identifying five different chromatin states (referenced by five colors) based on different combinations of chromatin component proteins. The RED and YELLOW chromatin correspond to active chromatin. The RED chromatin contains many tissue-specific genes and hotspots where many seemingly unrelated proteins co-localise. The YELLOW chromatin contains a majority of ubiquitously expressed housekeeping genes. The BLUE chromatin is characterized by the binding of Polycomb group proteins, which repress transcription. The BLACK chromatin is the most prevalent repressive chromatin type and contains two thirds of all silent genes. Finally, the GREEN chromatin is marked by the heterochromatin protein 1 (HP1) and SU(VAR)3-9, with several HP1-associated proteins and covers large domains in pericentric regions. However this state does not correspond to the repressive state usually attributed to the term heterochromatin but rather to a neutral alternative (reviewed in van Steensel, 2011). Furthermore, several recent papers have described more than two chromatin states by identifying combination of histone marks in different cell types or species (Kharchenko et al., 2010; Gerstein et al., 2010; Ernst and Kellis, 2010).

At a more detailed level of resolution, individual nucleosomes have been shown to displace in response to transcriptional activity. Nucleosome-free regions are often

observed upstream of the TSS and downstream of the 3'-end of expressed genes (essentially described in yeast, Mavrich et al., 2008). The upstream nucleosome-free region is likely to permit the assembly of the transcription machinery. More generally, the nucleosome-free region is created through the action of specific enzymes that replace histone molecules (SWR1 remodelling complex: H2A to H2A.Z), remove them (RSC complex and chaperones) or move them along the DNA (Swi/Snf complex). Therefore, nucleosome dynamics is important as it regulates DNA accessibility.

1.3.4.2 Histone variants and modifications

Isoforms of histone subunits and covalent modifications of the histone tails are important contributors for creating hetero- and euchromatin. In certain chromatin regions, nucleosomes may contain histone variants in place of a core histone subunit. This compositional difference confers special regulatory properties to those regions. Only variants for the core histones H2A and H3 are known so far. For example, H2A.Z and H3.3 histone variants are largely enriched in nucleosomes positioned close to the TSS of active genes (Malik and Henikoff, 2003), which together with covalent modifications (for example, methylation and acetylation) may help nucleosome eviction and pre-initiation complex assembly. H2AX and its phosphorylated form γH2AX are found at DNA-damaged sites and macroH2A 1 and 2 replace H2A on the inactive X-chromosome in female mammalian cells. Additionally, centromeres contain the histone variant CENP-A, which replaces H3 and is important for chromosomal segregation. Although very little is still known about their mechanism of action, these variants clearly play an important role in transcriptional regulation.

Histones can also be covalently modified. The three main types of modifications described so far are: Methylation (Me), Acetylation (Ac) and Phosphorylation (Ph). All these modifications have been shown to be involved in transcriptional regulation.

Additionally, they have been identified as playing a role in DNA repair, chromosome condensation and DNA replication (reviewed in Kouzarides, 2007). Such modifications are placed or removed by enzymes. These specific enzymes often function in large multi-protein complexes and can act on both histone and non-histone targets (Glozak et al., 2005). Importantly, these enzymes display remarkable specificity for particular target amino acid residues in the histone tails and work in well-defined cellular conditions depending on external and internal signals.

Numerous studies have been performed to discover proteins catalysing histone modifications and understand their mode of action. The major categories of known enzymes are histone acetyltransferases (HAT) for acetylation, kinases for phosphorylation, protein arginine methyltransferases (PRMT) and histone lysine methytranferases (HKMT) for arginine and lysine methylation respectively. Most modifications, including acetylation, phosphorylation and methylation marks are reversible since antagonist enzymes have been identified; these are histone deacetylases (HDACs), phosphatases (PPTases) and amine oxidases or hydroxylases. These enzymes are presented in Figure 1.7 along with their antagonist enzyme (if known).

Interestingly, some enzyme complexes have several catalytic sites and so perform multiple actions. For example NuRD which is recruited by Mi-2 in flies, combines ATP-dependent nucleosome displacement and deacetylase activities (Kehle, 1998), leading to transcriptional repression.

1.3.4.3 General mode of action of covalent histone modifications

Functional consequences of histone modifications can be seen in two main ways: establishment of global chromatin environments and the orchestration of DNA-based biological task. The first category refers to the role of histone modifications in hetero- and euchromatin formation and maintenance. The second refers to their importance in processes such as transcription, DNA repair, DNA replication and chromosome

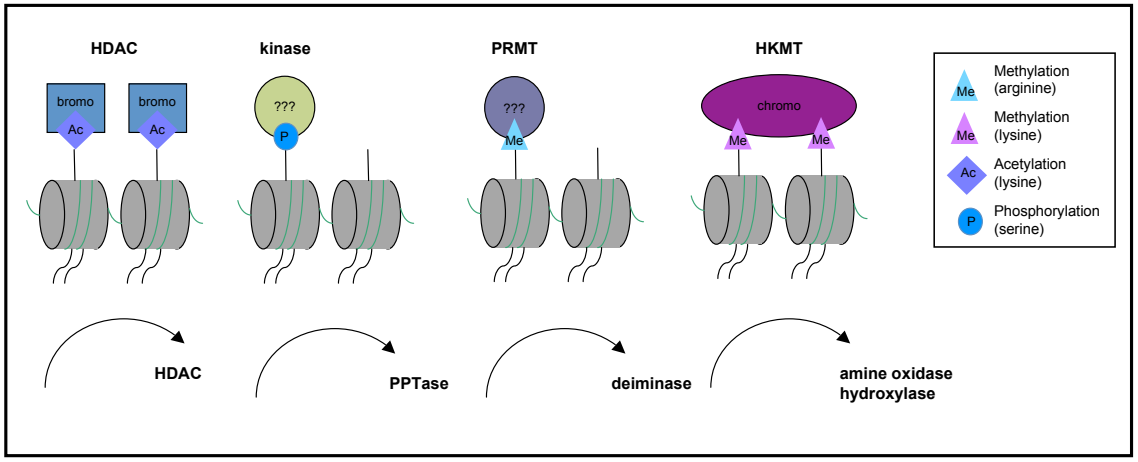


Figure 1.7: Histone modifying enzymes. The modifying enzymes are histone acetyltransferases (HAT) for acetylation, kinases for phosphorylation, protein arginine methyltransferases (PRMT) and histone lysine methytranferases (HKMT) for arginine and lysine methylation respectively. The major types of covalent histone modifications are transduced by histone-modifying enzymes (top) and removed by antagonizing activites (bottom). (Figure based on a figure from Allis et al., 2007)

condensation (Kouzarides, 2007).

How are these covalent modifications implicated in the transcriptional regulation? Modifications can directly alter chromatin structure: for example, acetylation is thought to relax chromatin and provide greater accessibility to the DNA by neutralising the positive charges found on lysine residues (Vettese-Dadey et al., 1996; Shogren-Knaak et al., 2006). Other modifications can help recruit specific regulatory proteins (in a context dependent manner); for example, chromo-like domains recognise methylations (Bannister et al., 2001; Kim et al., 2006), bromodomains recognise acetylations (reviewed in Bottomley, 2004) and a domain within 14-3-3 proteins recognises phosphorylations (Morrison, 2009).

1.3.4.4 Effects of histone covalent modifications on transcription

This section introduces the three major histone modifications and some of their known effects in greater detail.

Methylation has been linked to both gene activation and repression (Bannister and Kouzarides, 2005). Lysine methytransferases are very specific, acting on

one lysine on a single histone. The lysine-methylation marks linked to active transcription are H3K4me3, H3K36me3 and H3K79me. The latter two are implicated in transcriptional elongation (Carrozza et al., 2005). Methylation marks linked to transcriptional repression are H3K9me3, H3K27me3 (Roh et al., 2006). Similarly, arginine methylation has been implicated in positive and negative regulation of transcription.

Acetylated histone residues are almost invariably associated with active transcription; acetylation marks include H3K9ac, H3K18ac and H4K16ac. There are known transcriptional activators that recruit HATs to promote acetylation, and known repressors that recruit HDACs to reverse this modification.

Kinase activity has long been associated with signal transduction leading to changes in gene expression. More recently phosphorylation of H3S10 has been described as an important phosphorylation site for transcriptional regulation that is conserved from yeast to human. This example and others such as H3Y41 phosphorylation (Griffiths et al., 2010) shows that histone phosphorylation is likely to be important in transcription regulation.

In summary, active and inactive chromatin states are associated with distinct sets of modifications.

1.3.4.5 Combination of histone covalent modifications

Histone modifications do not occur in isolation but rather in a combinatorial manner. Multiple modifications can occur on residues within a particular histone tail and also in different histone tails or nucleosomes. Examples of the first are H3K9me/H3S10ph or H4S1ph, and H4R3me/H4K4ac. An example of the second is the ubiquitination of H2BK123 that leads to H3K4 tri-methylation. This in turn results in the acetylation of H3K9 and H4K16. Bromodomain-containing remodelling complexes then bind to these histones and displace nucleosomes to allow DNA-binding by TFs and Pol II

(reviewed in Allis et al., 2007)).

These observations gave rise to the concept of a histone code in which a single modification or groups of them have a predictable effect on transcription. Although this concept is useful to describe the need of a specific set of modifications for a particular task, it is unlikely to truly reflect the presence of a real code. Indeed some modifications are not present in all organisms, so the code, if it exists, would not be comprehensive and universal. Furthermore, since the chromatin can be modified by many remodelling complexes, the identification of general rules is difficult. Nevertheless, the availability of genome-scale techniques has considerably enhanced our ability to decipher patterns of some modifications. For example, there is a strong correlation between H3K4me3 and active promoter regions (Barski et al., 2007). Other modifications have different effects depending on the context. For example, H3K36 and H3K9 methylations have positive effects on transcription when they are found in coding regions and negative effects when in promoters (Vakoc et al., 2005). Bivalent domains in mouse embryonic stem cell (Bernstein et al., 2006) consist of genomic regions containing two methylation marks with opposing effects on transcription (H3K27me3 and H4K4me). These domains are present on low-expression genes coding for developmental TFs, suggesting that the modifications keep TFs poised for activation or repression until differentiation.

Finally, enzymes can also have different effects depending on the complex to which they belong. The histone demethylase LSD1 usually acts on H3K4 and represses transcription, whereas in a complex with the androgen receptor, it demethylates H3K9 and activates transcription.

1.3.5 Interplay between the transcriptional regulation levels

In order to regulate gene expression, there is a lot of interplay between the different levels of regulation that we discussed. Chromatin modifications are likely to impact

CRM activity by controlling accessibility to the DNA, and in return, binding sites can help recruit enzymes that affect chromatin structure. For example, the Twist TFs binds a set of CRM acting on genes involved in dorsoventral patterning and gastrulation during Drosophila embryo development. Then a bit later, it binds to different CRMs that regulate the expression of genes implicated in mesoderm migration and specification. The mechanism permitting this dynamic TF binding remains to be fully characterised but some scenarios are possible. One explanation is that inactive CRMs are within chromatin regions that prevent binding, which become exposed at later stages. Another example is nucleosome elimination from androgen receptor and FOXA1-binding sites in a human prostate cancer cell line in response to androgen stimulation (He et al., 2010). Finally, John et al. (2011) showed a dramatic dependence of the *de novo* genomic binding by the glucocorticoid receptor on preexisting chromatin accessibility.

1.4 Experimental techniques

1.4.1 Genome sequences and gene identification

To discover and understand genomic content and gene expression regulation, numerous methods have been developed. Functional genomic techniques such as microarrays that were developed for these purposes were in many cases developed by scaling up traditional molecular biological assays to the level of an entire genome.

One of the essential techniques is DNA sequencing. Sequencing techniques have developed hugely since the first methods were developed in early 1970s (Sanger and Coulson, 1975; Maxam and Gilbert, 1977). One of the major consequences of improvements to sequencing is the availability of full genome sequences. These genome sequences – particularly those of mammals – have generally arisen from

large international collaborations; for example, the Human Genome Project or the Chimpanzee Sequencing and Analysis Consortium (Lander et al., 2001; The Chimpanzee Sequencing and Analysis Consortium, 2005). The genomes for other models organisms such as yeast, worm and fly were completed in 1996, 1998 and 2000, respectively (Goffeau et al., 1996; Press et al., 1998; Adams, 2000). These data represent a milestone in our biological knowledge: they can directly offer new directions of research in evolutionary biology and are only the initial step in the functional genomic analyses.

Gene-finding algorithms (reviewed in Burge and Karlin, 1998) were developed to identify protein-coding genes in genomes. To understand the functions of these genes, high-throughput technologies such as RNAi-screens and microarrays were developed. Some of these methods have now been replaced by the arrival of Next Generation Sequencing (NGS) technologies.

1.4.2 Measurement of gene expression by microarray experiment

Microarray technology was one of the first high-throughput technologies developed to measure gene expression. This technology enables measurement of expression levels of thousands of genes at once. Two types of microarrays have mainly been used; oligonucleotide microarrays, containing short probes synthesised directly on the glass, and complementary DNA (cDNA) microarrays, consisting of longer probes that are deposited on the slide and are more commonly used for comparing hybridization between two RNA samples. For oligonucleotide microarrays, the probes are usually between 25 and 60 bases long. To measure gene expression levels, mRNAs are extracted from biological samples and fragmented. They are then labelled; for example in the Affymetrix platform they are reverse-transcribed into biotiny-

lated single-strand complementary DNA sequence. These are then hybridised to the corresponding microarray probes. The expression levels are measured using a laser-based scanner detecting the amount of fluorescence resulting from fluorescent streptavidin that bind to the biotinylated DNA sequences that have hybridised to the microarray. For cDNA microarrays, mRNA from two different biological samples are reverse-transcribed to cDNA labelled with two distinct fluorescent dyes. The cDNAs are hybridised to the microarray probes, and then scanned at different wavelengths to measure the amount of hybridisation from each sample.

Microarrays were first used to compare gene expression profiles of yeast in several cellular conditions (Spellman et al., 1998) and to identify abnormal gene expression in cancer samples (DeRisi et al., 1996). This technology was rapidly adopted for many different types of studies; for example to study the expression profile of disease samples (van 't Veer et al., 2002), the classification of samples according to their gene expression patterns (van de Vijver et al., 2002), identification of expression in intergenic regions (Bertone et al., 2004) and the determination of sequence variation between individuals of the same species (Jänne et al., 2004). An example of a major result from the technology is the accurate molecular cancer classification profiling of tumour gene expression (Ramaswamy et al., 2001), the identification of new gene variants and confirmation of genes playing a role in the development of seven common diseases such as diabetes and hypertension (Wellcome Trust Case Control Consortium, 2007). The microarrays are now very widely used.

1.4.3 Chromatin Immunoprecipitation followed by microarray hybridisation

Chromatin Immunoprecipitation (ChIP) is a commonly used technique to detect protein-DNA interactions. Formaldehyde treatment of cells allows cross-links to

form between the protein and DNA in vivo, thus stabilising protein-DNA complexes. The complexes are then extracted from lysed cells and sonicated to form DNA fragments that are about 200-600 base-pairs in size. The sample is immunoprecipitated using an antibody against the protein of interest (for example, a transcription factor or histone modification) to retrieve protein-DNA fragment complexes. The DNA fragments are dissociated from the protein and assayed by Southern blot or qPCR in order to identify the level of bound protein or modification at specific sites compared with a negative control such as genomic DNA or mock-IP.

This long-standing technique has been then coupled with microarray (ChIP-chip) to identify DNA-binding sites on a genomic scale. The immunoprecipitated DNA is hybridised to microarrays. For small genomes, all intergenic regions could be represented on the array, for larger genomes, such as human and mouse, ChIP-chip experiments were often done with whole-genome tiling sets of arrays to cover the entire genome. The computational analysis is thus similar to that for gene expression (section 1.4.2), except that a signal enrichment is expected only for immunoprecipitated samples. Bound genomic regions correspond to those that display significant enrichment in the microarray signal.

1.4.4 High-throughput sequencing applications

In the last few years, new expanded biological data sets have been generated through the development of the NGS technologies. In contrast to the semi-automated capillary-based methods previously used, NGS techniques use arrays or micro-scale beads as support for the material to be sequenced, which enables a much higher degree of parallelisation. The construction and amplification of the sequencing library are done in vitro. After library preparation, the PCR amplicon issued of any single library molecule end up spatially clustered on the array (or on the surface of a micro-scale beads, in function of the technique used). Sequencing by synthesis is performed

with alternating cycles of enzyme-driven biochemistry to produce the next base and imaging of the full array (reviewed in Shendure and Ji, 2008).

These new sequencing methods are faster and offer the possibility to sequence a large amount of nucleic acid at steadily decreasing costs. This new ability to sequence tens or hundreds of millions of short DNA fragments simultaneously allows generation of very large experimental datasets, at a resolution and accuracy that were unimaginable only a few years ago. NGS is now used in different areas of research such as whole genome sequencing, sequencing of mRNA for gene expression profiling (RNA-seq), detection of fusion genes from mRNA transcripts (cancer research), characterisation of structural variation, identification of DNase I hypersensitive sites, discovery of new small RNAs, determination of three-dimensional chromatin interactions (Hi-C), identification of DNA:protein binding events and detection of histone variants and modifications (ChIP-seq, Figure 1.8).

The new technologies have some disadvantages compared with microarray-based techniques. First is the higher cost, but this will improve over the next few years. Error rates are still relatively high, but again this is progressively improving with new updates to sequencing machines. On the whole however, they brought tremendous improvements to genome-scale experimental methods and they provide many advantages compared with microarrays. They offer full genome coverage, high resolution and less noise in general in comparison with previous techniques addressing the same questions. (reviewed in Park, 2009)

NGS technologies are gradually replacing microarrays in both expression and binding studies (RNA-seq and ChIP-seq). Both of these new methods have the advantages mentioned above compared to their microarray-based predecessors. First, RNA-seq experiments allow more accurate quantification of transcripts abundance; they provide greater dynamic range and with less noise than microarrays. Additionally this technique offers the possibility to detect new transcripts leading to the

identification of new alternative splice sites (Mortazavi et al., 2008; Wang et al., 2008; Sultan et al., 2008). Second, ChIP-seq experiments allow accurate identification of binding sites. This method is described in more detail below.

1.4.5 Chromatin Immunoprecipitation followed by high-throughput sequencing

The ChIP-chip method presented in section 1.4.3 is now almost entirely replaced by chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq). This method allows identification of DNA-binding sites at a genomic scale in an unbiased fashion as it does not depend on what is represented on the array. Additionally, it suffers from less noise because signal from cross-hybridisation is removed. An early study made by Johnson et al. (2007) shows the increased sensitivity and specificity of ChIP-seq genome-wide mapping of TF binding sites as well as the identification of non-canonical binding sites. Similarly, the first ChIP-seq studies of histone modifications suggested new functions for modifications and the importance of combinatorial modification patterns (Barski et al., 2007).

ChIP-seq data also requires extensive computational analysis. Briefly, sequencing data from IP samples are compared with data from a negative control such as genomic DNA to identify statistically significant binding sites. These sites then need to be interpreted in a biologically meaningful manner. Depending on the type of experiment performed, three types of signal can be retrieved: (i) localised binding detected as peak signals, (ii) localised but broader regions of binding extending up to a few kilobases, and (iii) broad regions of binding of up to several hundred kilobases. The first describes a classical signature of a TF binding to its cognate DNA sequence. The second can be observed for Pol II and some histone modifications that are confined to certain genomic loci such as coding regions. The last one refers to

histone marks such as H3K27me3 that extend across entire genomic regions that are repressed (Pepke et al., 2009). Therefore, different methods are needed to identify these distinct types of signals. As ChIP-seq data are still new, there is active development of new analysis methods. The first methods to be published were so-called peak-finding approaches such as ERANGE (Mortazavi et al., 2008), PeakSeq (Rozowsky et al., 2009), spp (Kharchenko et al., 2008) or MACS (Zhang et al., 2008). As datasets for histone modifications emerge, new methods for detecting other types of signals are progressively becoming available such as Repitools (Statham et al., 2010), CEAS (Shin et al., 2009), and EpiChIP (Hebenstreit et al., 2010).

1.4.6 Protein-protein interactions

The techniques presented so far aimed mainly at identifying gene expression and understanding its regulation. Other techniques have been developed to study the protein products of gene expression. Here, we will focus only on the ones dedicated to protein-protein interactions (PPI).

Proteins within cells generally interact with other proteins either transiently or as part of stable complexes. Several methods have been developed to identify PPIs. Yeast two-Hybrid (Y2H) is the main method for detecting direct, transient interactions. This method employs a transcriptional activator that has two domains: one that binds DNA, and another that coordinates the assembly of the pre-initiation complex. The bait protein of interest is expressed as a fusion with the DNA binding domain. The prey protein is expressed in a different yeast strain as a fusion with the activator domain. After mating of the two yeast strains, the expression of a reported gene is evaluated: if the proteins physically interact with each other, the DNA binding domain and activator domain are brought together and so activate the expression of a reporter gene.

Tandem Affinity Purification coupled to Mass Spectrometry (TAP-MS) and co-

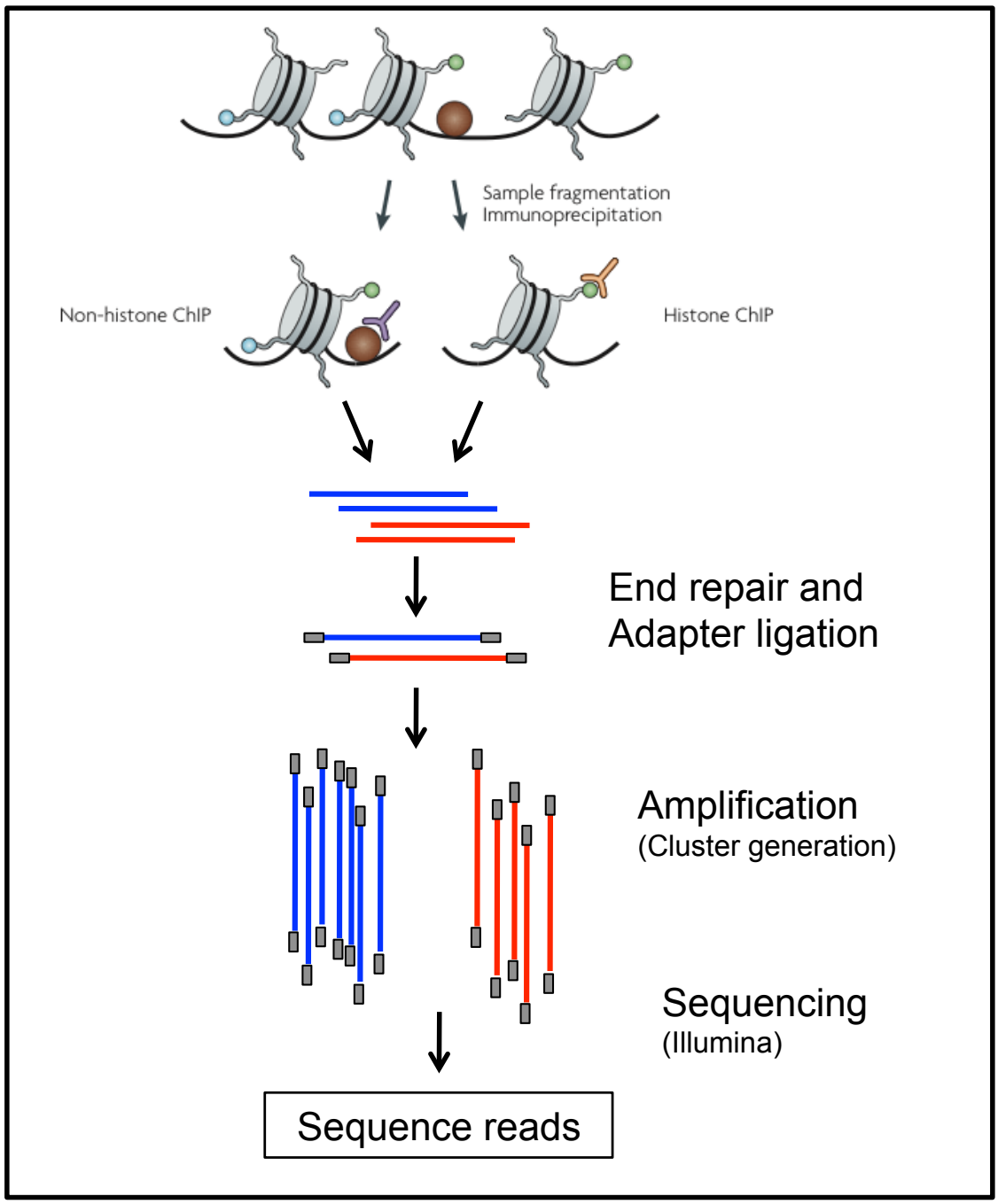


Figure 1.8: ChIP-seq expreriment workflow. ChIP-seq expreriment workflow. Chromatin immunoprecipitation followed by massively parallel sequencing. After immunoprecipitation of chromatin regions of interest, DNA is purified. Adapters are ligated to the DNA fragments, followed by an amplification step generating clusters on the flow cell. The clusters are then sequenced. Figure modified from Park, 2009.

immunoprecipitation (CoIP) are used to identify protein complexes. The TAP technique involves the fusion of a TAP tag to the protein of interest. The fused protein

is introduced into the host cell or organism and then allowed to interact with its normal partners. After cell extraction, the tagged protein and its partners are retrieved by affinity purification. The recovered complexes are then identified by mass spectrometry.

PPIs can also be confirmed and characterised in detail by methods such as X-ray crystallography and NMR, but these are still yet to be scaled up to a high-throughput level like the above techniques.

1.4.7 Public databases

The scientific community can access published datasets in public databases. The repositories are essential resources for the scientific community, as they provide access to very large amounts of data. Generally datasets are annotated to allow users to understand how they were generated, and also to offer easy searches. There are many different databases such as Ensembl (Hubbard et al., 2009), a database for eukaryotic genome sequence annotations, and Uniprot, a database of curated protein sequences.

In this thesis, we retrieved data from Ensembl for the identification of the fly and human genes. We used ArrayExpress (Parkinson et al., 2009), a database of functional genomic experiments including microarrays, to access human expression datasets. Finally, we also used a PPI dataset compiled from the HPRD, IntAct, BIND, DIP, and MINT PPI databases to analyse human protein interactions.

1.5 Computational techniques developed and used

Computational analysis methods are at the heart of bioinformatics work, allowing us to answer biological questions. For each experimental technique presented in the section 1.4 different analysis tools have been developed. Where possible, we

have used established methods, but sometimes it has been necessary to develop new methods when none were available.

The main steps of microarray analysis have become standardised in the past few years. The most common workflow for microarray data analysis involves: data acquisition; normalisation and pre-processing, unsupervised class discovery, detection of differential gene expression, which can lead to prediction, and functional interpretation of the results. However, this workflow can differ depending of the biological question being addressed. Further, some questions cannot be solved with the statistical methods developed so far. An example of this is presented in chapter 2 which describes a new method to detect condition- or tissue-specific gene expression.

This thesis also presents analyses of NGS datasets. As pointed out in the ChIP-seq computational analysis review by Pepke et al. (2009), several methods have recently been developed to detect and analyse peak binding signals present in ChIP-seq data. However broader signals such as Pol II-binding and histone modifications needed new algorithms. Chapter 4 adapts a method DESeq (Anders and Huber, 2010) that was originally developed for detecting differential gene expression for ChIP-seq experiments. Therefore, the thesis employs a combination of established and newly developed methods.

1.6 Aims of this thesis

The work presented in this thesis aims to increase our understanding of gene expression regulation through the analysis of genomic data using computational and bioinformatic approaches. Transcription regulation is assessed at two different levels: (i) at the TF level by the analysis of gene expression of different human tissues and the use of TF combinations and (ii) at the histone level by the analysis of the acetyltransferase MOF in *D. melanogaster*.

In order to perform this study we had the following objectives:

- to develop a statistical method for microarray data to identify genes that are specifically expressed in a small number of conditions;
- to identify the set of human genes and TFs that are specifically expressed in distinct tissues —an analysis of microarray data for 32 healthy human tissues;
- to identify groups of TFs that work together in different human tissues by integrating gene expression and PPI data;
- to identify targets of MOF-binding and its effect on histone acetylation in *D. melanogaster* using ChIP-seq experiments;
- to evaluate the effect of MOF on gene expression from microarray experiments;
- to investigate the effect of MOF on Pol II activity.

1.7 Chapter description

In the first part of this thesis, we investigate specific TF expression and their combinatorial action in humans. In chapter 2, we first present a new statistical method which detects, from microarray data, genes that are specifically expressed in a small number of conditions. We apply this method to an expression dataset for 32 healthy tissues and demonstrate its utility by identifying both known and new tissue-specific genes. In chapter 3, using a manually curated list of human TFs, we identify specifically expressed TFs in these tissues. Since TFs are known to work in a combinatorial manner, we build a network of co-regulating TFs from expression and protein-protein interactions data.

In the second part of this thesis, in collaboration with Dr A. Akhtar's experimental group at the Max-Planck Institute for Immunobiology and Epigenetics (Freiburg,

Germany), we investigate the effect of histone acetylation on gene expression using dosage compensation (DC) in *D. melanogaster* as a model. DC represents an excellent model for large-scale transcriptional regulation as it involves the up-regulation of the entire X chromosome in male flies. In chapter 4, we first identify target genes of MOF, a key histone acetyltransferase for DC. We analyse the pattern of H4K16ac and assess the effect that MOF has on gene expression. In chapter 5, we then investigate the effect that MOF has on Pol II function. We then propose a model for DC at the level of Pol II activity.

Finally, in chapter 6, we summarise the findings presented in the thesis, and propose further studies that will continue to increase our understanding of transcriptional regulation.

2

SpeCond: a method to detect condition-specific genes

2.1 Introduction

Cells in multicellular organisms share the same genomic information, but they express it in different ways to achieve cell-specific functions. Transcriptional regulation is the first step at which this specificity is determined, as it is the most basic level at which gene expression is controlled. Recent surveys of transcriptomics data across multiple cell types revealed two broad categories of gene expression: (i) ubiquitous; and (ii) tissue- or cell-type specific expression (Freilich et al., 2005; Vaquerizas et al., 2009). The first category contains genes that are used in most tissues at similar expression levels; they are thought to provide core cellular functionality (Warrington et al., 2000; Butte et al., 2001). The second category comprises genes with distinct expression in a small subset of tissues or conditions; these are likely to play a key role in defining cell-specificity.

Detecting specific expression can help to decipher the molecular basis of func-

tional differences between cells. Similarly, measurements of gene expression differences between healthy and diseased cell samples have contributed to our understanding of complex diseases such as cancer or Alzheimer's disease (Blalock et al., 2004). Tissue-specific genes might therefore be ideal candidates for disease diagnosis and treatment. Beyond gene function, the investigation of tissue-specific expression has been key to our understanding of transcriptional regulation at a basic level, such as the characterisation of mammalian promoters by analysing the promoter regions of genes detected as tissue-specific or not. A further example is the identification of functionally related tissues (Schug et al., 2005; Greco et al., 2008).

Genes displaying tissue- or condition-specific expression can be identified on the basis of their differential expression using traditional statistical approaches. In datasets with only a few conditions, it is possible to use approaches such as the standard or moderated t-tests (Smyth, 2005; Tusher et al., 2001; Zhang, 2007). However, this approach becomes difficult with increasing numbers of samples, as the number of pairwise comparisons multiplies. An alternative method is the non-standard ANOVA, which tests all possible groups of samples (for all possible group sizes) against each other. However, this involves computationally intensive dynamic programming and cannot detect specificity in a single condition as the variance cannot be estimated. Moreover, the method requires equal standard deviations between all groups being compared. This cannot be assumed as a general rule as genes might have similar expression levels for some conditions (and hence small standard deviations), and more divergent expression levels for others. A further alternative is the Tukey test. However, this method requires independence between groups and a normal distribution of group means, criteria which are not often fulfilled in microarray experiments. Importantly, most of these methods, as well as others, assume that expression values follow a single normal distribution. Here we demonstrate that this

assumption is generally not satisfied and that these methods do not model the data correctly, but rather lead to false positive results.

In order to tackle more specifically the problem of detecting condition-specific gene expression, two methods were recently developed. First, a method called Tissue Specific Genes Analysis (TSGA) (Chengyin, 2008) implements the ROKU approach (Kadota et al., 2006) based on Shannon entropy values, a measure of the overall tissue-specificity degree of a gene, followed by the outlier detection of Kadota *et al.* (Kadota et al., 2003). It returns a list of conditions in which each gene is specifically expressed. Unfortunately, this method depends on a set of ubiquitously expressed genes to model background expression levels—information that is generally not available prior to analysis. Furthermore, the TSGA method produces binary outputs—a gene is classified either as condition-specific or not, without ranking genes or conditions. This makes the resulting lists difficult to interpret biologically. Secondly, Vaquerizas *et al.* previously developed a propensity measure to detect tissue-specific transcription factor expression (Vaquerizas et al., 2009). This method returns the propensity for a given gene to be expressed at a certain level in particular conditions relative to its expression across other conditions. The method provides a good ranking of condition-specificity across samples. However, there is no control over the number of conditions in which a gene can be specific and there is no statistically meaningful threshold for specificity. As such, there is currently no straightforward, statistically robust method to detect condition-specific gene expression.

Here we present a new method called SpeCond to detect condition-specificity from a dataset of gene expression measurements. The method fits a normal mixture model to the expression profile of each gene, and identifies outlier conditions. We compare SpeCond against several alternative approaches using a gold stan-

dard dataset and demonstrate that SpeCond outperforms other methods. Finally we apply the SpeCond approach to a subset of the Genome Novartis Foundation SymAtlas dataset (Su et al., 2004), and identify specifically expressed genes from 32 human tissues samples which appeared meaningful after GO annotation analysis. The method is freely available as an R package within the Bioconductor software project (Ihaka and Gentleman, 1996; R Development Core Team, 2008; Gentleman et al., 2004, `http://www.bioconductor.org/help/bioc-views/release/bioc/html/SpeCond.html`).

It is worth noting, that in the literature, some authors make a further distinction between specific and selective genes: specific genes are significantly differentially expressed in only one tissue whereas selective ones are differentially expressed in a small group, with the precise number of tissues often being left to individual choice. For this work, we will refer to both categories as specific genes.

2.2 Methods: Detecting condition-specific gene expression

2.2.1 SpeCond in a nutshell

Briefly, SpeCond examines the distribution of expression values for each gene in turn. It identifies outliers that indicate unusually high or low expression in specific conditions relative to others. We define the background distribution—the distribution of expression values for a gene across conditions—using a normal mixture model. Using this background distribution, we compute a p-value for the expression of each gene value across all conditions. After repeating the procedure for every gene in the dataset, we correct all p-values for multiple testing. Finally, we identify condition-

specific expression values for each gene using a p-value threshold (Figure 2.1). The different steps implemented in the method are described in detail below.

2.2.2 Modelling the null distribution

Previous methods have modelled gene expression values using a Gaussian distribution. However, most datasets do not fit this distribution well, as they often exhibit varying degree of skewness. To overcome this, we use a mixture model that fits between one and three normal distributions to the expression profile of a given gene. This is achieved using the mclust package (Fraley and Raftery, 1999, 2003, 2006) in the R software environment. The algorithm performs a hierarchical clustering of a mixture model of normal distributions via Expectation-Maximisation (EM). The best fitting model is then selected using the Bayesian Information Criterium (BIC).

In order to define the null distribution from the fitting process of a given gene, we identify the mixture component(s) that correspond to outliers, i.e. condition-specific expression(s). First, we test whether the mixture component has a median value distinct enough from the median of the main component (see below). If this is true, we then evaluate the following two possible scenarios (Figure 2.2): (i) whether the mixture component represents a small proportion of the data and is well separated from the main component; and (ii) whether the mixture component represents a small proportion of the data and has a large standard deviation compared with the main component. Any mixture component that falls into any of these two scenarios is likely to contain specific expression patterns and will not therefore be included as part of the null distribution. Once all mixture components have been evaluated, the remaining components are combined using their means, standard deviations and relative weights. By default, if a single component fits the data, the mean and standard deviation from the initial parametrisation is used for the null distribution

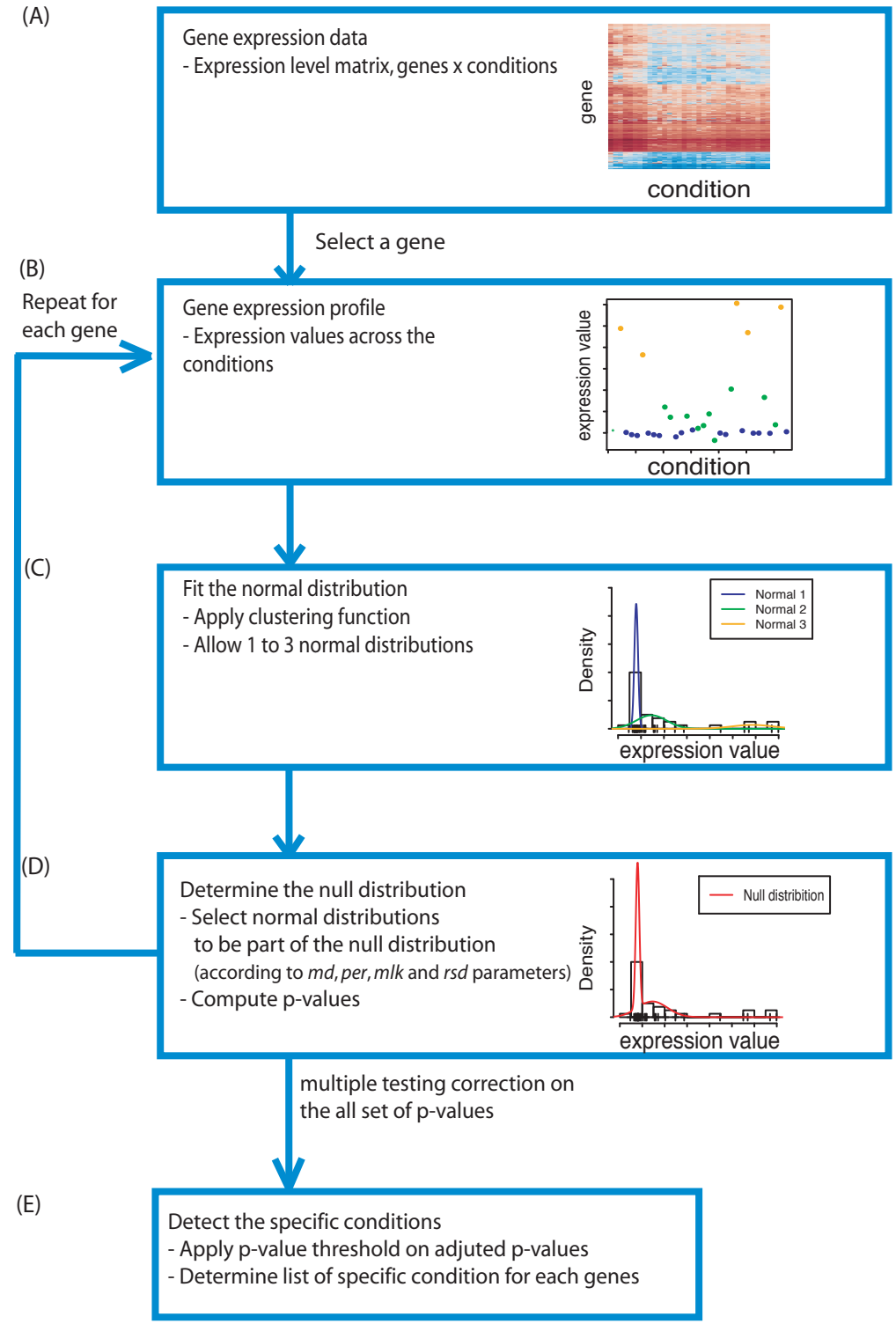


Figure 2.1: SpeCond workflow. From a gene expression microarray dataset (A) SpeCond uses a model of mixture of normal distributions (C) to determine the null distribution of every gene (D) and identifies the condition(s) in which the gene presents a statistically significant specific expression (E). The way the parameters *md*: (median difference), *per* (percentage of conditions that can be considered as specific), *mlk* (minimum likelihood) and *rsd* (ratio of standard deviation) are used to determine the null distribution is presented in Figure 2.2.

(Figure 2.1, D). As a result, our approach returns the optimal model for expression values after the identification of outliers.

2.2.3 Identifying condition-specific expression values

Next, we compute a p-value for every gene expression value to determine whether the gene is specifically expressed. These p-values are based on the null expression of each gene, and are computed as the sum of the p-value for each normal distribution weighted by their respective proportion. This procedure is applied independently for each gene. We then perform a multiple testing correction (Benjamini and Yekutieli, 2001) on the overall set of p-values to correct for the number of tests performed due to the number of genes. Finally, a gene is determined to be specific if at least one adjusted p-value is below the specified threshold (0.05 by default). As a result, SpeCond classifies each gene as either displaying specific expression or not and returns the list of condition(s) in which it is specific (Figure 2.1, E).

2.2.4 The tuning parameters

Three main sets of parameters control SpeCond's behaviour: (i) those controlling the implementation of the normal mixture model; (ii) those used to decide which normal distributions are used in the final null distribution; and (iii) a p-value threshold to define a gene as being condition-specific. Next, we will review these three sets of parameters in turn.

Firstly, two parameters (λ and β) control how gene expression data are modelled as a mixture of normal distributions. λ (defaults to 1) adjusts the weight of the effect of the number of parameters that need to be estimated when selecting the mixture model. As λ increases, the model penalises the inclusion of more parameters (ie, more normal distributions). β (defaults to 1) establishes a prior used to determine

the variance of the normal distribution (see SpeCond vignette).

Secondly, four parameters (*md*, *per*, *mlk* and *rsd*; Figure 2.2) control the inclusion of the different mixture components in the final null distribution. *md* (defaults to 0.75) is the minimum distance between median values of any pair of normal components. *per* (defaults to 0.3) is the maximum proportion of conditions in which a gene can be detected as specifically expressed. *mlk* (defaults to 25) is the minimum log-likelihood for a set of expression patterns given a pair of normal components. *rsd* (default to 0.1) is the ratio of the standard deviations of a pair of a normal components. A mixture component is excluded from the null distribution under the following circumstances: (i) when *md* is larger than the threshold and *per* and *rsd* are smaller than their respective thresholds—this scenario corresponds to normal components that are a small proportion of the data, separated from the main component and with large standard deviation; and (ii) when *per* is smaller than then threshold and *md* and *mlk* are larger than their respective thresholds—this scenario corresponds to normal components that are a small proportion of the data, well separated from the main component with little or non-existent overlap.

Finally, pv (defaults to 0.05) determines the adjusted p-value threshold under which genes are classified as condition-specific. It is worth noting that in order to consider multiple gene expression profiles, SpeCond uses a two-step selection procedure. In the first step SpeCond runs with a stringent set of parameters (step 1: $\lambda = 1$, $\beta = 6$, $md = 0.75$, $mlk = 5$, $rsd = 0.1$, $per = 0.1$ and pv = 0.05), aimed to detect condition-specific expression for genes with significantly high expression value (often happening in a single condition). In the second step, SpeCond uses a stricter set of parameters (step 2: $\lambda = 1$, $\beta = 0$, $md = 0.75$, $mlk = 25$, $rsd = 0.1$, $per = 0.3$ and pv = 0.05) to allow detection at a finer level of detail. A full description of the parameters can be found in the user guide document of the SpeCond package in Bioconductor (see Appendix A).

2.3 Detecting tissue-specificity across the human genome

We used SpeCond on the SymAtlas dataset (Su et al., 2004) that contains genome-wide expression profiles for 79 human tissues and cell lines. To avoid tissue redundancy, which would affect the selection of tissue-specific genes, we focused on 32 major healthy tissues and organs present in the dataset (Table A.1). We performed quality checks using the arrayQualityMetrics package (Kauffmann et al., 2009) and processed the raw data using the three-step GCRMA algorithm as implemented in the Bioconductor project (Wu et al., 2004). We then computed the mean of the log2 expression values of the two replicates for each tissue and used it as an expression value for a given probe set in a given tissue. We utilised the annotation available in the Ensembl database (Ensembl v52) to map 17,064 probesets to 11,713 Ensembl ID genes. Only probe sets that mapped uniquely to genes were used for this analysis. We applied SpeCond to this dataset with the following parameters: (step 1: $\lambda = 1$, $\beta = 6$, $md = 0.75$, $mlk = 5$, $rsd = 0.1$, $per = 0.1$), step 2: $\lambda = 1$, $\beta = 0$, $md = 0.75$, $mlk = 25$, $rsd = 0.1$, $per = 0.3$ and pv = 0.05 (Figure 3). Finally, in order to consider a gene as tissue-specific, we required that all associated probe sets were detected as specific in a given tissue. This resulted in the detection of 2,673 human genes identified as specific for at least one but no more than nine tissues. The combination of parameters was chosen to achieve the best sensitivity at a 5% false positive rate as measured using Receiver Operating Characteristic (ROC) curves (Figure 2.4).

Figure 2.3 depicts a heatmap representation of the gene expression profiles for

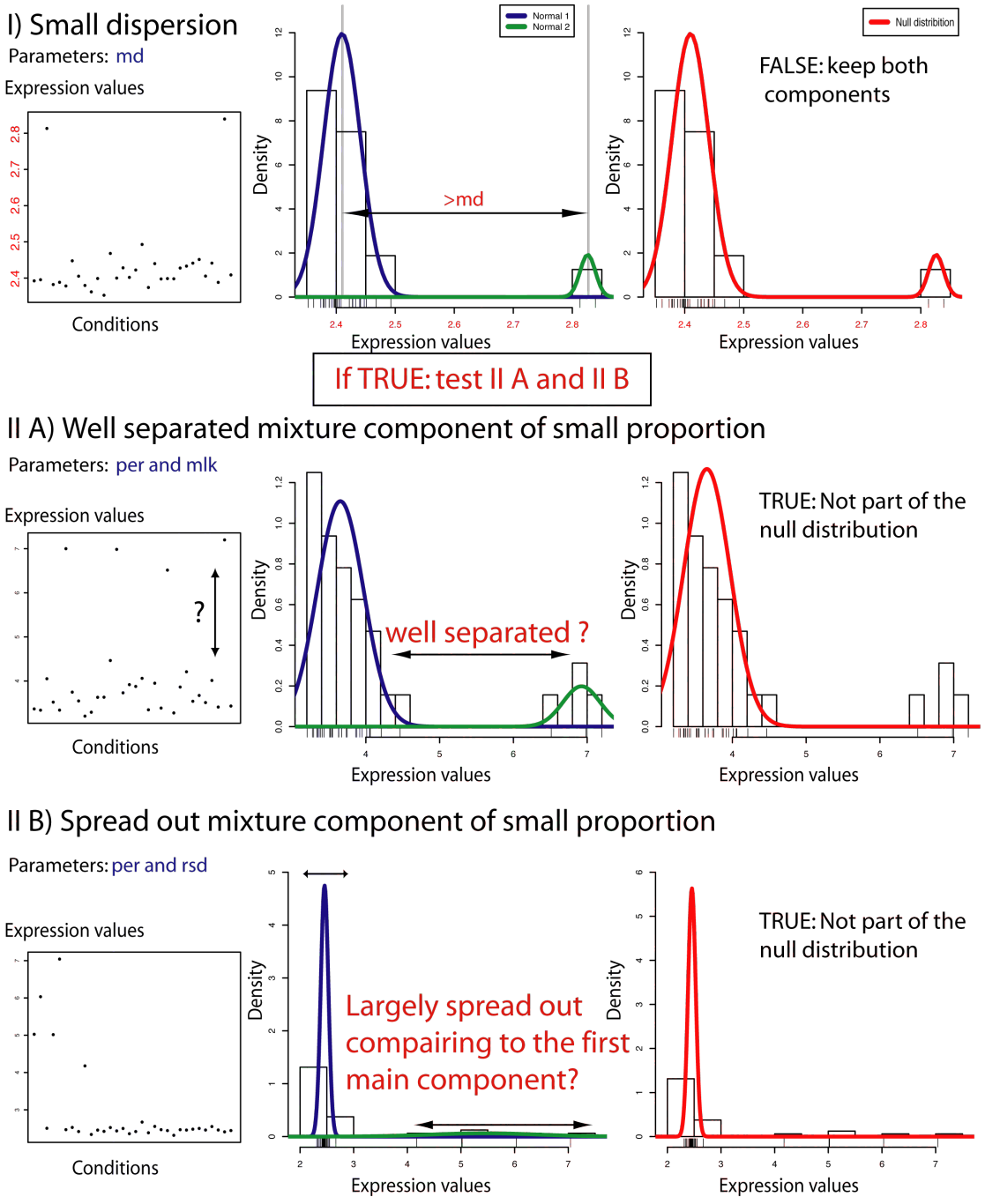


Figure 2.2: Determination of the null distribution. Three different conditions are evaluated in order to consider a normal component as part of the null distribution. (i) The median of the values from each component must have a difference larger than *md*. If this first condition is fulfilled, the procedure tests the following conditions: the normal component will not be part of the null distribution if: (ii) the normal component is small and well separated, i.e. the minimum of the absolute log-likelihood ratio of the expression values under the two components is larger than *mlk*; or (iii) the normal component is small and largely spread out, i.e. the standard deviation ratio is smaller than *rsd*.

tissue-specific genes across the 32 tissues analysed. Among these we can observe 1,133 genes that are specific in just one the tissues examined. These genes could be considered tissue markers: such as Y box binding protein 2 (YBX2) only detected in the testis, which is known to be a germ-cell specific gene implicated in the regulation of the stability and/or translation of germ cell mRNAs (Tekur et al., 1999) or the COBL-like 1 (CPBLL1) gene specifically expressed in the placenta but not functionally characterised so far. The remaining 1,540 genes share specificity among several tissues. This is particularly noticeable in sets of related tissues, or those that share a common ancestor. For example, tissues from the nervous system—brain, foetal brain and spinal cord—share 144 genes with neural-related specific expression patterns.

To assess the biological significance of the results obtained with SpeCond, we performed a Gene Ontology (GO) enrichment analysis for each set of tissue-specific genes using g:Profiler (Reimand et al., 2007). For 28 out of the 32 analysed tissues, the molecular functions and pathways detected as significant corresponded to what we expected to be prominent in the given tissue. For example the GO terms "contractile fiber" and "heart morphogenesis" are enriched in heart, "spermatogenesis" is specifically enriched in testis, and "T cell activation" is enriched in the thymus. To measure the global performance of SpeCond, we computed a general log-score for each tissue as the sum of the logarithms of significant p-values of the individual GO enrichments and compared it with log-scores obtained from random sets of genes. For all tissues, SpeCond log-scores were significantly better than those of random sets of genes. The large majority ($\approx 99\%$) of genes that were specific were due to an up-regulation in a few tissues (1 to 9 tissues). Tissue-specificity due to down-regulation of gene expression, i.e. tissue-specific repression, was detected very rarely across the dataset.

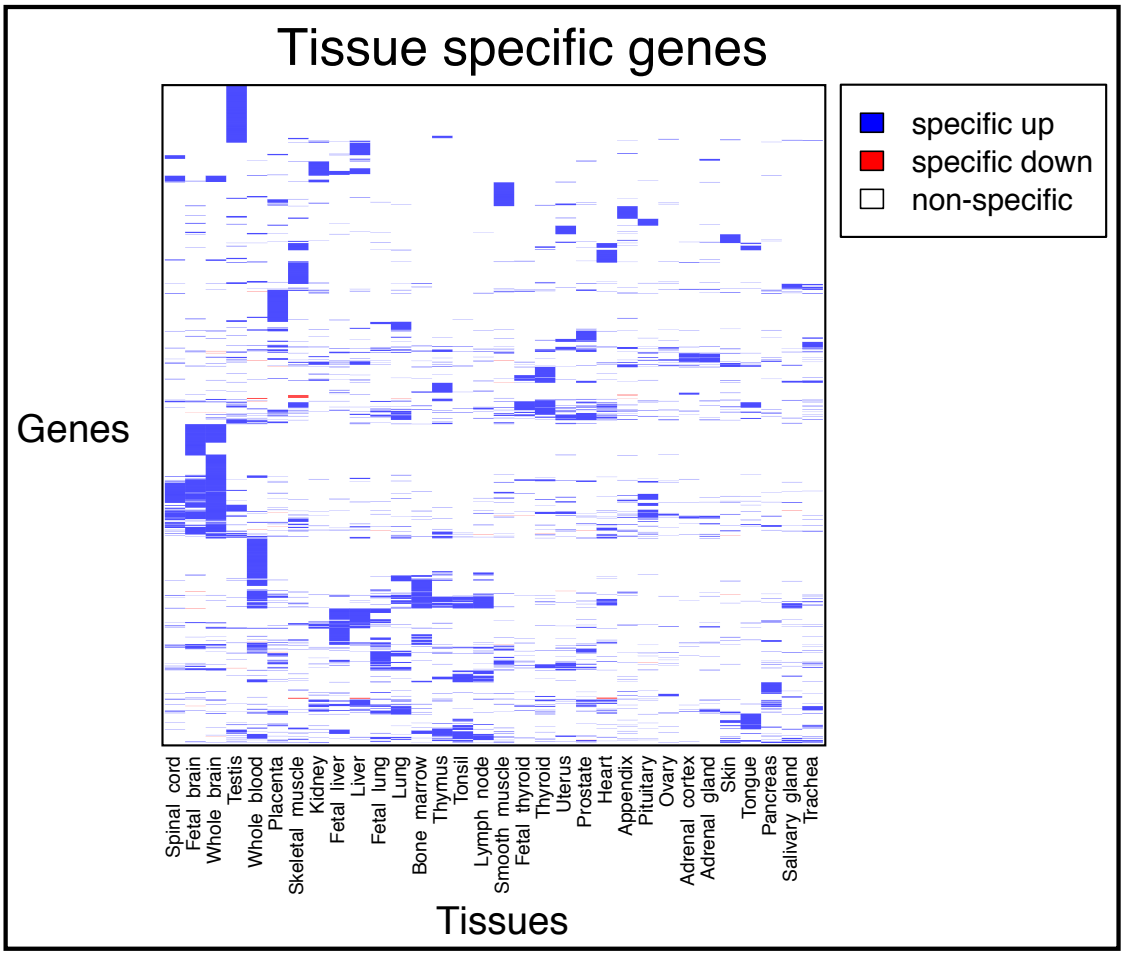


Figure 2.3: Heatmap representation of the tissue-specificity of each gene. The specific behaviour of every specific gene (y axis) in every tissue (x axis) are represented by a coloured box; blue if the gene is specifically up-regulated in the tissue, red if the gene is specifically down-regulated in the tissue, and light grey if the gene does not present any specific expression for the tissue.

Next, we focused on exploring individual tissues. As an example, a closer examination of the 287 liver-specific genes detected by SpeCond showed many genes that are important for liver functions, such as amino and fatty acid metabolic processes or gluconeogenesis. Among this group of genes, we found examples of genes previously known to have liver-specific expression, such as NR1I3, a key regulator of xenobiotic and endobiotic metabolism (Yamamoto et al., 2003), INSIG1 which takes part in metabolic control (Peng et al., 1997) or SDS, an enzyme involved in

metabolising serine and glycine (Sun et al., 2005a; Yamada et al., 2008). In addition, we found genes that had not been assigned to have a liver-specific function. One example is ATF5, implicated in differentiation, proliferation and survival in different cell types but whose liver function has not yet been assigned. ATF5 is detected by SpeCond as specifically highly expressed in the liver and very recently, the first studies of its function as a regulator of the hepatic stress response have just been performed (Pascual et al., 2008). Turning to other tissues, we can highlight that the transcription factor TTF1 (Homeobox protein Nkx-2.1, Thyroid transcription factor 1) is detected as specific in the fetal thyroid, thyroid and in the lung. This transcription factor gene is currently understood as binding to and activating the promoter of thyroid-specific genes as well as being essential for morphogenesis and differentiation of thyroid, lung and ventral cortex (Boggaram, 2009).

Another example is illustrated in organs of the central nervous system. These tissues—brain, foetal brain and spinal cord—present the largest list of tissue-specific genes (511 for brain, 406 for foetal brain and 266 for spinal cord). Functional profiling of tissue-specific genes shared by the three tissues revealed well-known nervous-tissue functions such as "generation of neuron", "axonogenesis", "synaptic transmission", as well as the neural cellular component "neurofilament cytoskeleton". In addition, we were able to identify EAAT1 (Excitatory amino acid transporter 1), SLC1A3, Sodium-dependent glutamate/aspartate transporter 1) as specific in the three tissues outlined above. This gene is known as a member of a family of high-affinity sodium-dependent transporter molecules that regulate neurotransmitter concentrations at the excitatory glutamatergic synapses of the mammalian central nervous system (Kirschner et al., 1994). Further, we detected many genes with expression profiles specific for these tissues that have not previously been experimentally associated with any neural function (Table A.1 in Appendix A). Among

these we found ZNF365 and ZNF536, two transcription factors previously reported to have brain- and spinal cord-specific expression (Vaquerizas et al., 2009).

2.4 Comparison with other approaches

2.4.1 Gold standards

We evaluated the accuracy of SpeCond by comparing the results obtained with our method against other available approaches, namely TSGA and propensity. We performed this by computing ROC curves using a reference set of gold standard ubiquitous and tissue-specific genes (Figure 2.4). For tissue-specific genes—our positive control—we computed the intersection of human tissue-specific genes from the Tissue-specific Gene Expression and Regulation (TiGER) database (based on EST data in 30 human tissues Liu et al., 2008) and those determined by Dezso and colleagues (expressed in one tissue among the 31 analysed in microarray expression analysis (Dezso et al., 2008)). Overall, our positive control group contained 3,984 probe sets in a total of 26 tissues, resulting in 5,595 gene-specific conditions. In order to obtain a set of ubiquitously expressed genes—our negative control—we merged two datasets: (i) the list of negative strand matching probe sets found in Affymetrix's HG-U133a array (Warren et al., 2007); and (ii) the union of housekeeping genes detected in two independent studies (Eisenberg and Levanon, 2003; Dezso et al., 2008). In total our negative control contained 3,657 probe sets. Since these probe sets correspond either to genomic loci that should produce background signal -negative stand matching probe sets as they do not correspond to gene coding regions- or to genes that do not display tissue specificity across 30 human tissues, we are confident that these lists represent an unbiased estimate of ubiquitous expression. However, it is worth noting that due to differences in origin, handing

and sample preparation, the overlap and detection of tissue-specificity might not be exactly the same across all datasets.

2.4.2 ROC curves

We first tested the effect of different SpeCond parameters (Figure 2.4, A, B). This allowed us to determine the combination of parameters that performed best for the GNF dataset. Next we computed ROC curves for the GNF dataset and compared the specificity and sensitivity of SpeCond against that obtained by the TSGA method and by the propensity approach as implemented in Vaquerizas et al. (2009). As neither of these methods is able to detect tissue-specific repression, only genes that showed tissue-specificity for expression were considered in this analysis. To perform a valid comparison of each method, we established a common threshold for false positive detection at 5% and adjusted the parameters of each method to provide the best sensitivity (number of true positives). Using SpeCond, we were able to obtain the best sensitivity (62%) of all methods, with an error rate of 5% (Figure 2.4, A). TSGA also showed good performance, with a sensitivity of 60% at a 5% error rate. SpeCond showed a significant improvement in sensitivity when compared with the propensity method (62% compared to 55% at a 5% error rate). Furthermore the false positive rate of the propensity method increases rapidly without an improvement in sensitivity. Finally, we performed GO enrichment analysis for the tissue-specific gene-sets returned by TSGA and the propensity method. We computed a general log-score as before to compare the performance of each method from a biological perspective. The log-scores obtained were 18,315.89, 17,663.82, and 15,629.51 for the SpeCond, TSGA and propensity method, respectively. SpeCond and TSGA showed similar enrichment levels and a great improvement from the propensity method. Overall therefore, SpeCond displays better sensitivity and specificity levels than any of the other available methods.

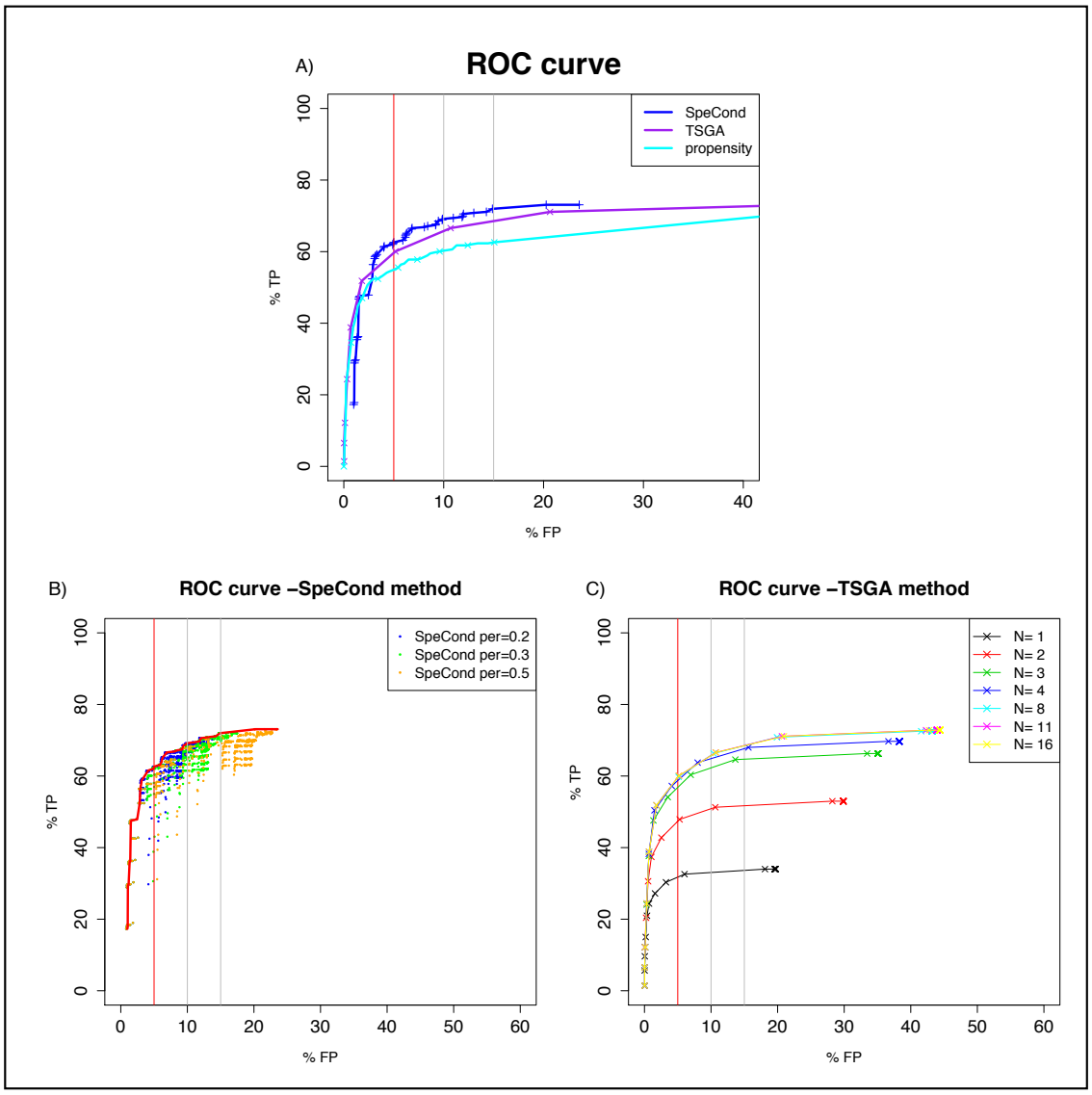


Figure 2.4: ROC curves. Evaluation of the performance of the three methods: A) SpeCond (blue), TSGA (purple), Propensity (cyan). Detailed parameter evaluation for the SpeCond and TSGA methods: B) SpeCond: the set of dots corresponds to the results obtained for different parameters sets (see below) using *per*=0.1 in step 1 and three *per* parameter values in step 2; *per*=0.2 (blue), *per*=0.3 (green) and *per*=0.5 (orange), respectively. With the above mentioned *per* values, we used β=6 and 0 in step 1 and step 2, respectively and varied in the two steps the parameters *mlk* (from 0 to 300) and *rsd* (from 0 to 2). The red curve represents optimal parameter set for each FP rate. C) TSGA for the maximum number of specific tissue that can be detected N=1, 2, 3, 4, 8, 11, 16, varying the H.critical value from 1 to 19.8.

2.5 Bioconductor R package

In order to provide easy access to the method, we developed SpeCond, an R package integrated within the Bioconductor software (freely available from `http://www.bioconductor.org/help/bioc-views/release/bioc/html/SpeCond.html`).

The input to the software package is a matrix of normalised expression values in which rows correspond to genes or probe sets, and columns correspond to different conditions. The package returns different outputs: (i) R objects, (ii) text files and (iii) HTML pages. The user can thus further use these R objects and text files for further analysis. Additionally, a general HTML results page provides an overall view of the condition-specific behaviour for the entire dataset (Figure 2.5, and 2.6). This results page contains information about the total number of specific genes by condition, an expression specificity heatmap and a table containing the names and numbers of condition-specific tissues for each specific gene. An individual results page can also be generated for each gene (Figure 2.7, Figure A.1 in Appendix A) in which SpeCond displays the expression profile for the gene, the density curves for the mixture model fit, the null distribution, the set of parameters used for the analysis and finally, the list of condition(s) in which a gene is specific with the associated adjusted p-values. The large set of visualisation functions for the results enable the user to perform a full analysis. He/she can test different configurations of the parameters to evaluate which combination correctly corresponds to its particular dataset.

AGC_SpeCond_results_ENSG_cons...

Tissue specific results for human genes (ENSG conservative)

Parameter set used to detect the specific condition(s)

	beta	lambda	per	md	mlk	rsd	pv
Step 1	6	1	0.1	0.75	5	0.1	0.05
Step 2	0	1	0.3	0.75	25	0.1	0.05

Table 1: The parameter set used for the analysis

Specific genes - conditions

# conditions	1	2	3	4	5	6	7	8	9	-	sum
# genes	1133	487	310	163	170	145	97	107	61	-	2673

Table 2: Number of genes detected as specific in the different number of conditions

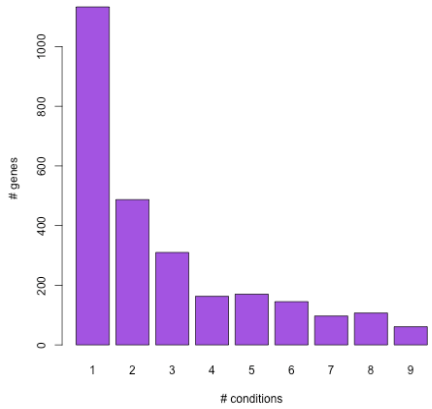


Figure 1: Barplot representing the number of specific genes (y axis) in function of the number of their specific condition (x axis)

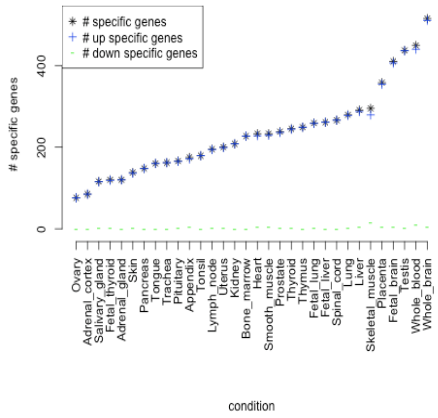


Figure 2: The number of specific genes detected in each of the conditions

Figure 2.5: SpeCond general HTML output (part 1). The HTML page displays a set of tables and figures; the parmeters used in the analysis (Table 1), the number of gene detected as specific in different numbers of conditions (Table 2 and Figure 1), and the numbers of specific genes (up-, down- and total) detected in each condition (Figure 2).

Heatmap representation

Profiles; 0,1 or -1 values

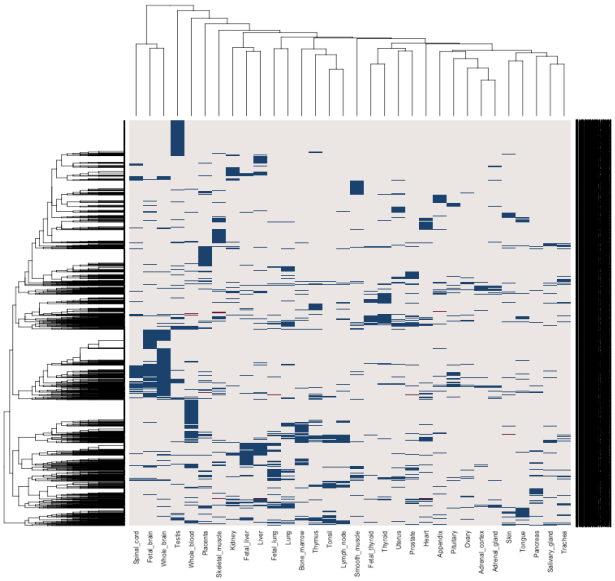


Figure 3: Profile heatmap representing the gene specificity (y axis) over the conditions (x axis), the colors correspond to the specific level; red, white, blue: down (-1), non (0), up (1) specific or just non (0) and specific (1) (red and blue respectively) if only two colors are present

Specific genes table result

Gene	Type	# condition specific	# condition up	# condition down	Condition up	Condition down	Gene page*
ENSG00000211973	S	7	7	0	Lymph_node,Lung,Thyroid,Tonsil,Bone_marrow,Thymus,Trachea	-	NA
ENSG00000198671	S	5	5	0	Whole_brain,Liver,Skeletal_muscle,Fetal_liver,Kidney	-	NA
ENSG00000170468	S	2	2	0	Testis,Thymus	-	NA
ENSG00000153287	S	1	1	0	Testis	-	NA
ENSG00000219491	S	1	1	0	Thyroid	-	NA
ENSG00000211663	S	6	6	0	Lymph_node,Heart,Tonsil,Bone_marrow,Salivary_gland,Trachea	-	NA
ENSG00000211639	S	1	1	0	Lymph_node	-	NA
ENSG00000216786	S	8	8	0	Adrenal_cortex,Appendix,Fetal_thyroid,Ovary,Skeletal_muscle,Tongue,Salivary_gland,Skin	-	NA
ENSG00000149397	S	2	2	0	Bone_marrow,Fetal_liver	-	NA
ENSG00000163918	S	4	4	0	Tonsil,Fetal_liver,Testis,Thymus	-	NA
ENSG00000211889	S	5	5	0	Lymph_node,Tonsil,Whole_blood,Bone_marrow,Thymus	-	NA
ENSG00000211829	S	2	2	0	Whole_blood,Thymus	-	NA
ENSG00000215430	S	1	1	0	Testis	-	NA
ENSG00000211598	S	7	7	0	Lymph_node,Heart,Tonsil,Bone_marrow,Salivary_gland,Thymus,Trachea	-	NA
ENSG00000211949	S	8	8	0	Lymph_node,Heart,Lung,Thyroid,Tonsil,Bone_marrow,Thymus,Trachea	-	NA
ENSG00000211940	S	9	9	0	Lymph_node,Heart,Lung,Thyroid,Tonsil,Bone_marrow,Salivary_gland,Thymus,Trachea	-	NA
ENSG00000211898	S	8	8	0	Lymph_node,Heart,Lung,Tonsil,Bone_marrow,Salivary_gland,Thymus,Trachea	-	NA
ENSG00000162722	S	2	2	0	Bone_marrow,Fetal_liver	-	NA
ENSG00000182901	S	2	2	0	Fetal_brain,Whole_brain	-	NA
ENSG00000198626	S	1	1	0	Heart	-	NA
ENSG00000187730	S	1	1	0	Whole_brain	-	NA
ENSG00000008130	S	3	2	1	Whole_blood,Bone_marrow	Uterus	NA
ENSG00000142733	S	1	1	0	Lung	-	NA

Table 3: Specific gene result: list of all the specific genes detected, the number of condition in which they are specific (all, up and down) and the respective condition names

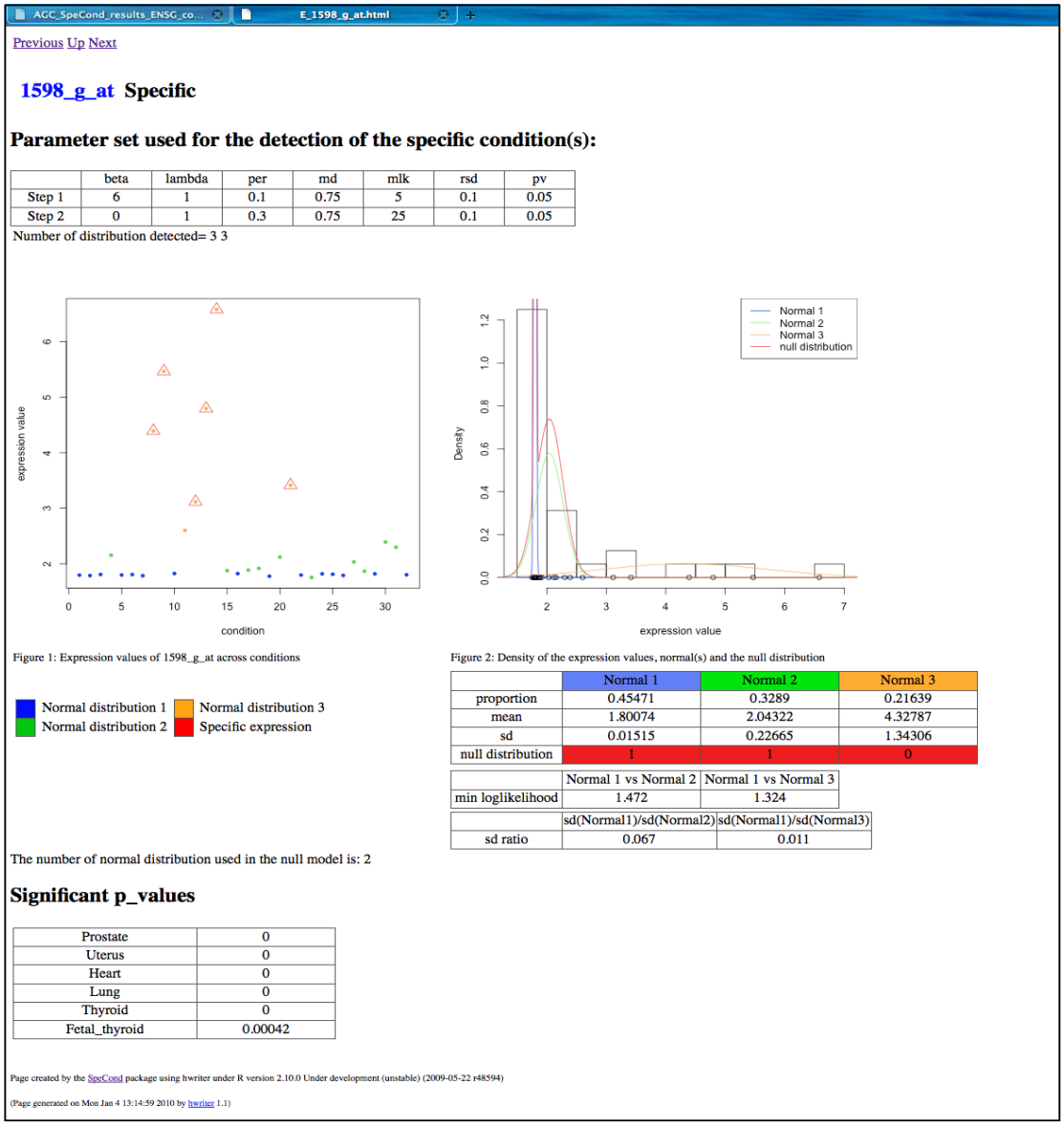
*, NA ???: You can generate the gene result page(s) of the gene present in the above table with the "getGeneHtmlPage()" function. Using its result object as the "gene.page.info" argument of the "getFullHtmlSpeCondResult()" function (function generating this page), you will have a direct link to the gene pages

The corresponding file is : AGC_General_Result/AGC_result_specific_probeset.txt

Page created by the SpeCond package using hwriter under R version 2.10.0 Under development (unstable) (2009-05-22 r48594)

(Page generated on Mon Jan 4 13:21:16 2010 by hwriter 1.1)

Figure 2.6: SpeCond general HTML output (part 2). Following the tables and figures presented in Figure 2.5, the SpeCond general HTML page presents a heatmap of the tissue-specific genes (Figure 3) followed by a table containing all the tissue-specific genes with the numbers and tissues in which they are detected (up and down are separated). Links to an individual SpeCond specific HTML page such as Figure 2.7 are present in the rightmost column of the table if the pages have been previously generated by the user.



AGC_SpeCond_results_ENSG_co... E_1598_g_at.html

Previous Up Next

1598_g_at Specific

Parameter set used for the detection of the specific condition(s):

	beta	lambda	per	md	mlk	rsd	pv
Step 1	6	1	0.1	0.75	5	0.1	0.05
Step 2	0	1	0.3	0.75	25	0.1	0.05

Number of distribution detected= 3 3

Figure 1: Expression values of 1598_g_at across conditions

Figure 2: Density of the expression values, normal(s) and the null distribution

Normal distribution 1
Normal distribution 2
Normal distribution 3
Specific expression

	Normal 1	Normal 2	Normal 3
proportion	0.45471	0.3289	0.21639
mean	1.80074	2.04322	4.32787
sd	0.01515	0.22665	1.34306
null distribution	1	1	0

	Normal 1 vs Normal 2	Normal 1 vs Normal 3
min loglikelihood	1.472	1.324

	sd(Normal1)/sd(Normal2)	sd(Normal1)/sd(Normal3)
sd ratio	0.067	0.011

The number of normal distribution used in the null model is: 2

Significant p_values

Prostate	0
Uterus	0
Heart	0
Lung	0
Thyroid	0
Fetal_thyroid	0.00042

Page created by the SpeCond package using hwriter under R version 2.10.0 Under development (unstable) (2009-05-22 r48594)

(Page generated on Mon Jan 4 13:14:59 2010 by hwriter 1.1)

Figure 2.7: Individual SpeCond specific HTML page; output for a specific probe set. Example of the 1598_g_at probe set detected as specific in 6 tissues. The HTML displays the probe set (or gene) name and a set of tables and figures: the parameters used in the analysis (top table), the expression profile (Figure 1), and the density curves of the mixture model fitting the expression values (Normals 1, 2 and 3, in blue, green and yellow respectively), as well as the null distribution (red) (Figure 2). The parameters of each normal distribution as well as the SpeCond parameter values are presented in the table below the Figure 2. Finally, the tissues in which the gene is detected as specific with their corresponding adjusted p-value are presented in the bottom table.

2.6 Discussion and conclusion

The widespread use of microarrays in biological research over the past few years has generated a flood of data characterising gene expression across many tissues in different species (Lukk et al., 2010). Determining tissue- or condition-specific expression from these datasets is an important aspect of genomic analysis, but it is difficult to perform using traditional statistical techniques. Here we have presented SpeCond, a new, statistically method to detect condition-specific expression from microarray data. SpeCond does not impose a single normal distribution to estimate the underlying distribution but computes an estimate of the null distribution using a normal mixture model. SpeCond is an ideal choice when no previous data about the organisation of the system under study are known as it is not assumed that the measured expression values follow a single normal distribution. As proof of principle, we have analysed the popular GNF SymAtlas dataset using SpeCond. The results showed a significant enrichment of tissue-specific GO annotations across genes detected as tissue-specific.

We evaluated SpeCond's performance against different alternative approaches to detect tissue-specific expression by comparing the sensitivity and specificity of tissue-specific detection against a gold standard. In all cases, SpeCond displayed higher sensitivity and a lower false discovery rate when compared with alternative approaches. Further, the enrichment of tissue-specific GO annotations was also superior for SpeCond compared to other methods. Importantly, the SpeCond package is not a black box. The user is encouraged to test different parameter sets to find the best sets returning meaningful results according to relevant biological questions. Indeed, the large set of visualisation tools allows the user to examine expression patterns in detail, to verify the fitting of the mixture normal distribution, as well as to compare easily the overall specific gene sets resulting from the use of different

sets of parameters.

A further advantage of SpeCond is its ability to generate ranked lists of genes based on their tissue-specific expression. The ability to classify genes regarding their contribution to tissue-specificity should be helpful for experimentalists that wish to identify candidate genes for detailed follow-up studies. In addition, these ranked lists can be used in computational approaches, such as the examination of the organisation of tissue-specific transcriptional networks or the putative annotation of unknown gene functions based on their expression pattern.

It is important to note that the number of conditions present in a dataset has an important effect on the results. Due to the intrinsic nature of the analysis, enough conditions need to be measured to determine whether tissue-specific expression is biologically relevant, i.e. if a dataset does not cover enough conditions, it will not be clear whether a gene is truly only expressed in a single tissue. Moreover, due to the null distribution estimation, the detection is more powerful and robust as long as the number of conditions increases. Therefore, we recommend that researchers use SpeCond with at least ten conditions in order to obtain relevant results. In future, it will be very interesting to analyse RNA-seq data with the same purpose. However, the model will need to be modified as the normal distributions do not fit count datasets. A negative binomial distribution as used in the DESeq method (Anders and Huber, 2010) is certainly more appropriate, and therefore a mixture of negative binomial distribution model would need to be created. Furthermore, currently available datasets are still restricted to a small number of conditions for which this method is not appropriate.

Finally, SpeCond is immediately applicable to many datasets measuring gene expression, including the detection of tissue-specific alternative splicing, in any species.

3

Human tissue-specific expression, a TF view

3.1 Introduction

Transcriptional regulation is the most basic mechanism that cells use to control the production of the right amount of protein. This regulation is mediated by a concerted effort that involves transcription factors, co-factors and chromatin-remodelling complexes working in a coordinated fashion. This in turn recruits the RNA polymerase and associated factors to promote or repress gene expression. Sequence-specific DNA-binding transcription factors (TFs) play a crucial role in this mechanism by recognising specific DNA sequences in gene promoters or enhancers and regulating the recruitment of the transcription machinery at the TSS. They are therefore key factors in establishing specific expression patterns across different cell types. Our group recently published a manually curated compendium of 1,391 human TFs in collaboration with Dr Sarah Teichmann's laboratory (Vaquerizas et al., 2009). This chapter presents an analysis of how these TFs are utilised in different tissues, in order to understand how they might regulate tissue-specific expression.

Please note that this work was started in parallel with a recent publication by Ravasi *et al.*, and unfortunately there are many overlaps in our findings (Ravasi et al., 2010).

Given the large number of target genes and vast number of conditions to which cells must adapt, the human genome does not encode for sufficient TFs to provide this regulation if they acted individually. It is therefore believed that most transcriptional regulation by TFs is achieved through their combinatorial action. Though combinatorial TF usage has been shown for individual examples, our understanding of the global impact of this mode of regulation is still limited. The first evidence of combinatorial regulation came from observations that many TFs interact with each other as homo- or heterodimers (reviewed in Luscombe et al., 2000). A prominent example is for instance the human TF Fos which participates in complexes to regulate cellular proliferation, transformation and death (Shaulian and Karin, 2002). In general, the combination of two TFs allows the recognition of a composite binding site that can be used to define a new set of specific targets. Furthermore, TFs have been identified in large complexes such as the enhanceosome that often include a variety of co-factors and are powerful transcriptional activators. Additional evidence comes from recent studies revealing the concerted action of several TFs binding to *cis*-regulatory modules (CRMs) in *D. melanogaster*. CRMs act as modular units that integrate the input from multiple TF-binding events to produce specific spatio-temporal gene-expression patterns. These studies also showed that the absence of a single factor can have a profound effect on the regulatory activity of other combinatorial TF partners (Bonn and Furlong, 2008).

A recent study by the FANTOM consortium presented the first large-scale analysis describing the combinatorial logic of TF interactions in a large number of human and mouse tissues (Ravasi et al., 2010). This integrative work demonstrated the importance of TF combinations for cell fate determination. For instance, they identified a set of six TF-TF interactions that best discriminate human tissues according

to their embryonic origin as well as a TF complex that operates in monocyte development. Ravasi and colleagues produced the most complete and up-to-date collection of protein-protein interactions (PPIs) between TFs identified by mammalian two-hybrid experiments.

In this chapter, we present a computational analysis of combinatorial TF function based on gene expression and PPI data. The availability of datasets describing the expression of the TFs in healthy tissues presented in the previous chapter, and the descriptions of physical interactions between them offer the possibility to examine the combinatorial TF network underlying tissue-specificity. To this end, we first identified all expressed and specifically expressed TFs in each of the studied tissues. We then examined the protein-protein interactions between TFs, showing that there is an enrichment of intra-family interactions. Coupling PPI information and TF expression, we detected the TF-TF interactions that are likely to occur in different tissues, which serves as an indicator of combinatorial regulation. Finally, we showed that adding the PPI information strongly increases our power to explain the downstream gene expression patterns of different tissues.

3.2 Data and Methods

3.2.1 Transcription factor expression analyses

Our analysis was based on the repertoire of sequence-specific DNA-binding TFs in the human genome, originally published by Vaquerizas et al. (2009). The dataset contained 1,391 manually curated human TFs from 41 different TF families (Vaquerizas et al., 2009). Here, we only considered the 23 families containing more than 5 TF members, in order to assess intra- and inter-family PPIs.

To analyse the expression level of these TFs in human healthy tissues, we used

the publicly available SymAtlas dataset introduced in chapter 2 (Su et al., 2004). This dataset contains genome-wide expression profiles for 79 human tissues and cell lines obtained using Affymetrix microarrays. To avoid redundancy in tissue-types, which would affect the selection of tissue-specific genes, we focused on a subset of 32 major healthy tissues and organs. Microarray data quality control, normalisation and further processing to obtain the log2 expression values of the probe sets were performed as described in (chapter 2, section 2.3).

To identify the set of expressed genes in each tissue, we applied the "Presence-Absence calls with Negative Probe sets" (PANP) method as implemented in R (Warren et al., 2007). This method makes use of hybridisation signals from negative-strand matching probe sets present in the microarray to model the background signal. As reported in Vaquerizas *et al.*, this method achieved better sensitivity and specificity compared with more common approaches like MAS5.0. We flagged a gene as expressed if hybridisation signals from all uniquely mapping probe sets exceeded the PANP threshold across all replicate arrays.

To identify tissue-specific TFs in each tissue, we applied SpeCond on the GNF dataset as described before (chapter 2, section 2.3). Since not all human genes are represented on the Affymetrix U133A GeneChip, we obtained expression information for 877 TFs.

3.2.2 Protein-protein interactions

We used a recently updated PPI dataset presented in Ravasi et al. (2010). In this study first, all PPIs between TFs were collected from the open-access databases HPRD, IntAct, BIND, DIP and MINT, and only interactions identified by low-throughput molecular techniques were retained. Next these data were complemented by PPIs identified using a mammalian-2-hybrid (M2Y) screen. The combination of the two datasets resulted in a list of 5,238 PPIs; of these 1,441 PPIs were between

546 different entries in our repertoire of DNA-binding TFs.

3.3 Transcription factor expression in 32 human tissues

First we examined the pattern of TF expression across the 32 human tissues. We identified 545 TFs (62% of the TFs on the microarray) that are expressed in at least one tissue. As previously reported, we observed a bimodal distribution of TFs that are expressed in only a few tissues, others that are expressed across all tissues, and very few TFs that are expressed in intermediate numbers of tissues (Vaquerizas et al., 2009). The same trend is observed for the expression of all genes (Freilich et al., 2005).

To define the set of specifically expressed TFs, we use the result of the SpeCond analysis. 146 TFs were identified as displaying tissue-specificity, with TFs being expressed in between one and eight tissues. Within each tissue, there were between 3 and 26 specifically expressed TFs. It is important to note that SpeCond not only detects TFs that are expressed in only a handful of tissues, but also those that might be broadly expressed but displaying specific over-expression in a few tissues. We also detected four TFs that were specifically under-expressed in some tissues, but they were not included in the analysis.

Among the specific TFs were some that were previously known to play a major role in their respective tissues. For example, NR1I3 was flagged as specific in liver and the TF is known to regulate xenobiotic and endobiotic metabolism (Yamamoto et al., 2003). In another case, TTF1 (Homeobox protein Nkx-2.1, Thyroid transcription factor 1) was flagged as specific in the fetal and adult thyroids, and in the lung; this TF is understood to bind and activate transcription at the promoters of

thyroid-specific genes, as well as being needed for morphogenesis and differentiation of thyroid, lung and ventral cortex tissues (Boggaram, 2009).

We also observed a large proportion of tissue-specific TFs that have not been characterised so far. For example ZNF365 and ZNF536 were detected in the whole brain and spinal cord, but there are no annotations indicating their function in these regions. Such examples suggest that the patterns of tissue-specific expression might help to inform future studies about possible regulatory functions.

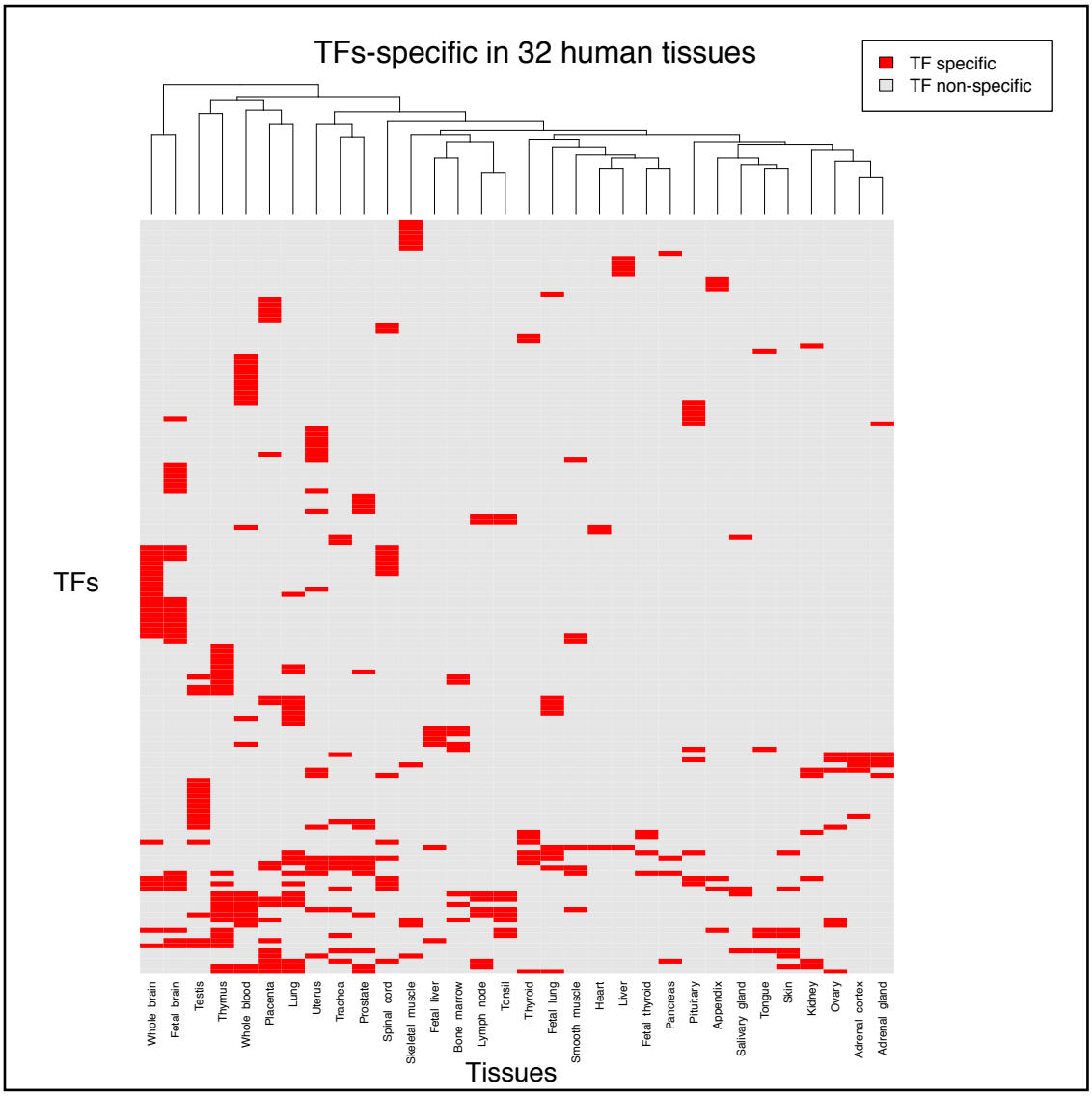


Figure 3.1: Heatmap of the tissue-specific TFs.

Using the information about the presence or absence of TFs in the 32 tissues, and

their specificity as defined by SpeCond, we classified TFs into three categories: (i) specific; (ii) intermediate; and (iii) general. Specific TFs are those defined as tissue-specific by SpeCond. General TFs are those that are not detected by SpeCond and expressed in at least 27 tissues. The remaining TFs were classified as intermediate. Each category was populated with roughly one-third of the 545 TFs in the dataset (Table 3.1).

TF classes	# (%)
General	173 (33%)
Intermediate	206 (39%)
Specific	146 (28%)

Table 3.1: Numbers of TFs in the three TF classes. Specific TFs are detected as specific in at least one tissue (up to 8) using SpeCond. General TFs correspond to those TFs that are not tissue-specific and expressed in at least 27 tissues. Any TF that is expressed and does not belong to any of these two classes is classified as intermediate.

In summary, this analysis provided an overview of how TFs are expressed across 32 human tissues. Given that there are many known key regulators for different tissue types, it is clear that specific TFs are important for establishing tissue-specific expression patterns. Of interest are those specifically expressed TFs whose regulatory function remains uncharacterised, since they are likely to play an important role in the tissue concerned. However, since TFs are also known to function in a combinatorial manner, the presence of a TF alone is unlikely to be sufficient to explain all tissue-specific expression. Therefore, next we proceeded to investigate the PPIs between TFs. PPIs contain TF pairs that have been identified as interacting directly or as part of a complex. This type of evidence is not exhaustive enough to obtain all real TF sets that work together in a given tissue, but allow us to go a step further toward the identification of TFs that are likely to function together.

3.4 Human TF-TF protein-protein interactions

We used the recent dataset of PPIs reported by Ravasi and colleagues to examine the physical interactions between TFs.

3.4.1 PPIs characteristics

The dataset contained 1,441 PPIs between 546 TFs (39% of the TFs list) including 175 (12%) homodimers. Interactions were defined as direct or "potentially indirect" depending on the experimental method that was used to identify them. Direct PPIs refer to experimental methods that can only detect proteins that are in direct, physical contact with each other, such as yeast-2-hybrid. Indirect PPIs refer to methods such as TAP-tagging followed by mass spectrometry that identify members of a stable complex; though proteins may also contact each other directly, we did not have enough information to make this distinction.

We identified 616 (43%) direct and 825 (57%) potentially indirect interactions between different TFs. The entire set of PPIs represents just 0.15% of all possible interactions between 1,391 TFs. Although TFs are unlikely to be so promiscuous in their interactions with each other, this dataset probably underestimates the real number of TF-TF interactions. Indeed Ravasi *et al.* reported 25% sensitivity for their M2Y experiments. Nonetheless, this dataset is the most complete and up-to-date collection of PPIs for TFs available so far, and it is likely to provide a suitable representation of TF-TF interactions.

3.4.2 PPIs and TF families

Given the importance of homo- and heterodimer formation to TF function, we evaluated the level of PPIs across different TF families. TFs are commonly classified into families according to the DNA-binding domain that they contain. This classifi-

cation has proved to be useful, as some TF functions can be predicted according to family membership (for instance, homeodomains are well-known for their regulation of developmental processes), and as it allows the study of TF evolution. Some TF families are well known for dimer formation; examples include members of the bZIP and helix-loop-helix families. Dimerisation constrains the evolution of the protein-interaction surface, the DNA-binding domain and associated target sites, since it imposes that the two members of the dimers "co-evolved" to continue to be able to interact and work together. Further, the two close recognition sites on which they bind need to remain targets of the two members of the dimers.

We evaluated the number of PPIs comprising TFs from the same or different families. Here, we excluded families containing five or fewer members. A PPI was classified as intra-family if the pair of TFs had at least one DNA-binding domain-type in common. As expected considering the relatively small size of the PPI dataset, we identified a significant enrichment of intra-family interactions 656 intra-family PPIs (46%, p-value$<$ 2.2 10^{-16}, Fisher's test). Only five families were not enriched for intra-family interactions (High-mobility group, zinc finger-GATA, Forkhead, POU, and AP2). 785 (54%) PPIs were classified as inter-family, but no family was enriched for inter-family interactions (Figure 3.2).

Next we evaluated the presence of direct or indirect PPIs. We found a significant enrichment of direct intra-family PPIs (p-value$=$ 1.62 10^{-3}, Fisher's test). One might think that the intra-family enrichment is essentially due to homologous heterodimers, however we recovered the same results after removing these dimers from the analysis. The fact that our results show enrichment of intra-family interactions for the very large majority of the TF families reflects that in general TFs interact more with other TFs from the same family.

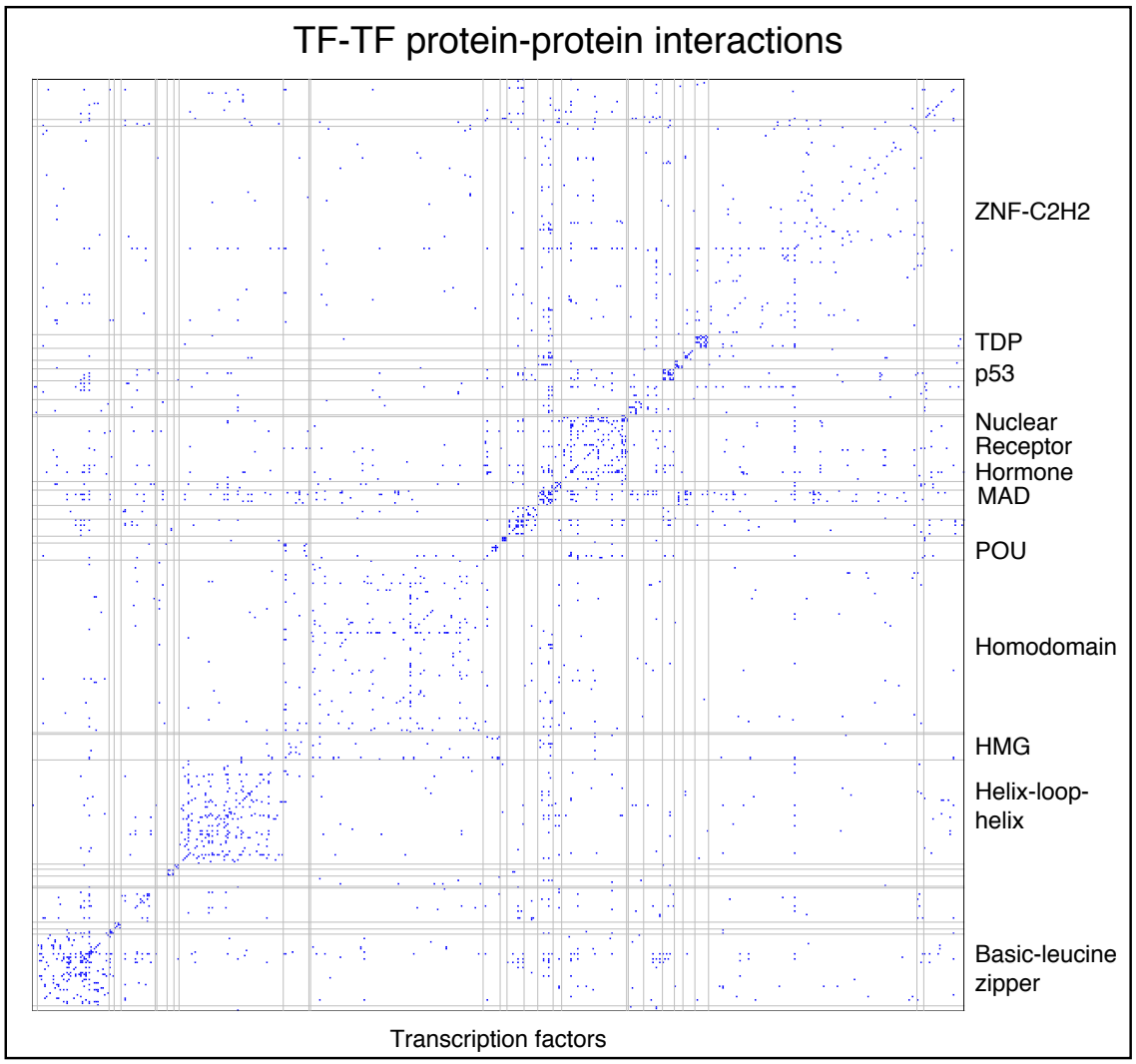


Figure 3.2: Heatmap display of human TF-TF interactions. TFs for each family are displayed on both axes in the same order. Blue points indicate a reported interaction between two TFs. Grey lines mark the boundaries between TF families. Names of major families are indicated on the left (ZNF-C2H2: Zinc finger - C2H2-type, TDP: transcription factor E2F/dimerisation partner, HMG: high-mobility group).

3.4.3 Connectivity and TF classes

We next proceeded to evalute the connectivities of TFs in our dataset. Figure 3.3, A shows the distribution of the number of interactions displayed by a given TF. We observed a characteristic power-law distribution, in which many TFs have a few PPIs (155 have only one interacting partner) and a few TFs act as hubs with many interactions (11 TFs have more than 26 interacting partner).

Next, we computed the mean number of PPI for the TFs in each of the TF classes, namely general, intermediate and specific (see section 3.3). These classes reflect characteristics of the TF expression that are expected to be linked to the their role in the regulatory network. We observed that specific TFs are involved in fewer interactions than intermediate and general TFs (Figure 3.3, B). This is in agreement with recent results reported by Ravasi *et al.* and adds further support to the notion that specific TFs take part only in a few different complexes to confer tissue-specific regulation. This is in contrast to general TFs, which are able to interact with many others; such TFs would be able to take part in many different complexes, so taking on a variety of different regulatory functions depending on the context.

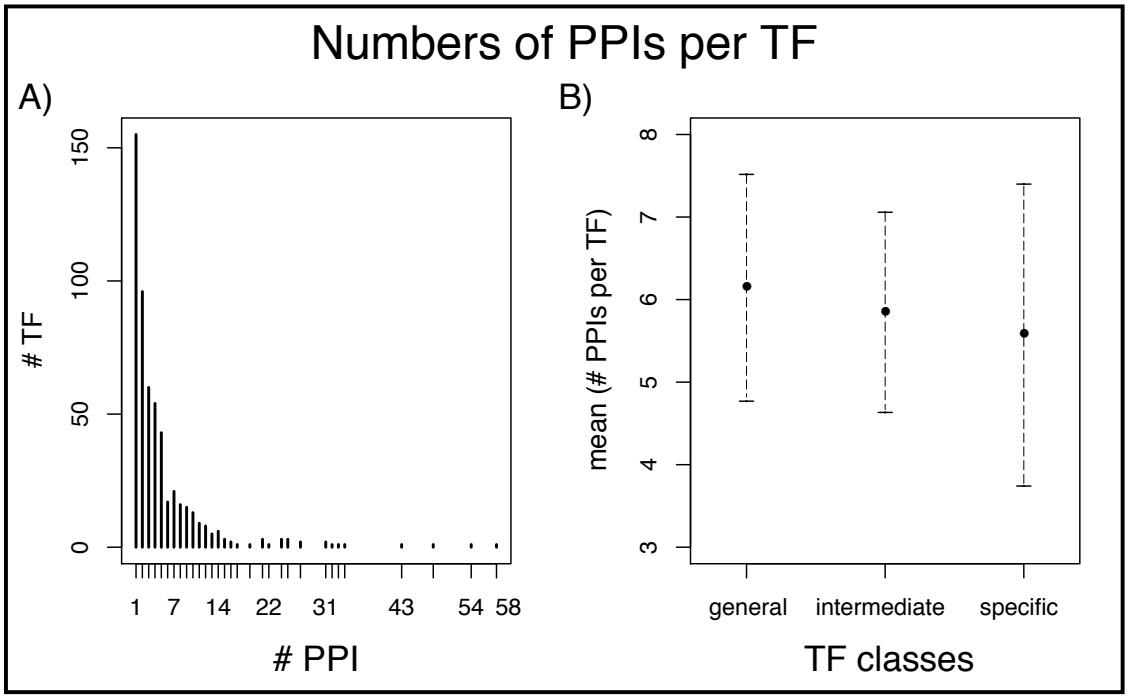


Figure 3.3: TF connectivity. A) Number of PPIs per TF. B) Mean numbers of PPIs for TFs in the three different expression classes: general (TFs expressed in at least 27 tissues), specific (TFs defined as tissue-specific by the SpeCond method) and intermediate (the remaining TFs). While the numbers of PPIs per TF tend to be higher for the general TFs as compared to the other classes, the difference is not statistically significant (p=0.73 and 0.15, Wilcoxon test performed between the general and intermediate sets and the intermediate and specific sets, respectively).

3.4.4 TF-TF protein-protein interactions in human tissues

Next we explored the presence and absence of different TF-TF pairs across different tissues. For each PPI, we identified the tissues in which they might potentially occur as those in which both TFs were expressed. Among the 1,190 PPIs for which expression data were available, we found 624 PPIs that were present in at least one tissue.

Figure 3.4 shows a distribution of the number of tissues in which different PPIs are present; as for the individual TFs, the distribution follows a bimodal distribution in which PPIs are present either in very few or almost all tissues. In a similar manner

to individual TFs, we classified these TF-TF pairs into three groups according to their expression: (i) specific; (ii) intermediate; and (iii) general. Specific PPIs are those present in fewer than five tissues, general PPIs are those detected in at least 28 tissues, and intermediate PPIs are the remaining ones (Table 3.2).

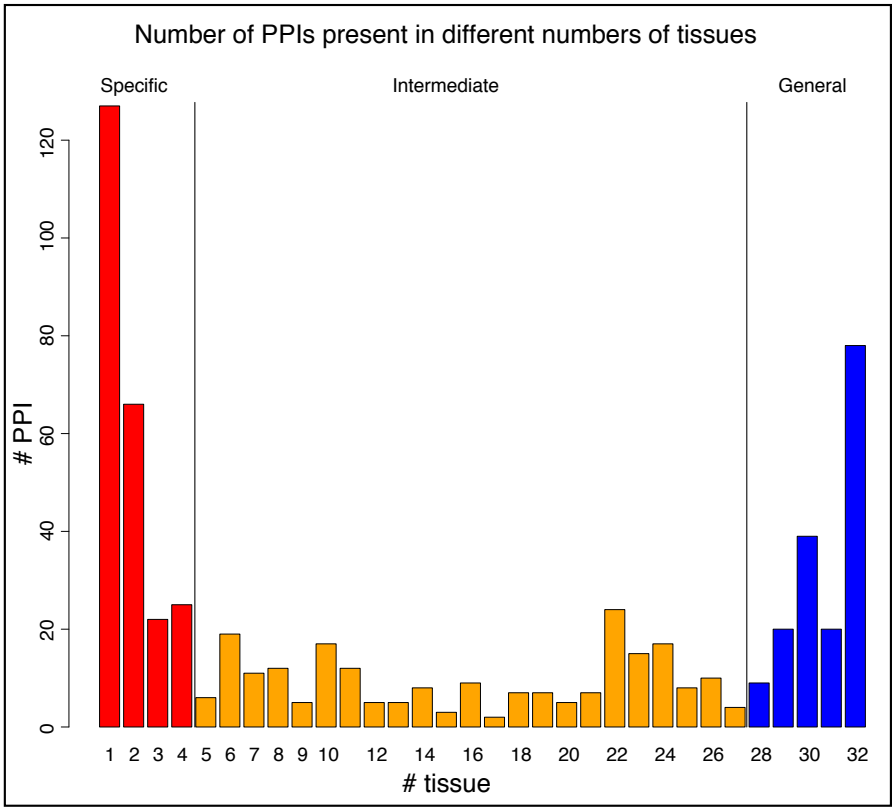


Figure 3.4: Number of PPIs per tissue. The three PPI classes based on their presence in the 32 tissues are indicated: general (blue), intermediate (orange), and specific (red).

PPI classes	# (%)	range (min-max)
General	166 (27%)	118 - 166
Intermediate	218 (35%)	30 - 187
Specific	240 (38%)	0 - 59
Total	624	153 - 393

Table 3.2: PPI classification. Number of PPIs for each expression category. PPIs in the general class are present in at least 28 tissues, those in the intermediate class are present in 5 to 27 tissues, and the ones in the specific class are present in a maximum of 4 tissues.

In line with our earlier observations for the expression of individual TFs, we suggest that PPIs appearing in only a few tissues are particularly interesting because they are more likely to play an important role in defining tissue-specific expression.

We first examined the PPIs that we defined as specific, and assessed the contribution of individual TFs: are specific PPIs composed of specifically expressed TFs, or can non-specific TFs come together to form a specific PPI? The 425 specific PPIs involved 193 different TFs (each PPI was counted in every tissue it was specifically expressed). As expected, we observed that these PPIs were significantly enriched for specifically expressed TFs (p-value$<$ 2.2 10^{-16}, Fisher's exact test). Further, PPIs in 18 out of the 32 tissues displayed significant contributions from individually specific TFs. Therefore the specificity of many TF-TF pairs is already defined by the expression pattern of the individual TFs.

However interestingly, we found that an even larger proportion of specific PPIs (256 PPIs; 60%) consisted of TFs that were not themselves specifically expressed. In other words, the specificity of these PPIs result from pairs of TFs that are individually expressed in a large number of tissues, but coincide only in a small number of tissues (Figure 3.5). We cannot rule out the possibility that two TFs present in a given tissue and known to interact do not actually interact in this tissue, since another co-factor or a particular phosphorylation state of the protein(s) might be missing to allow the interaction. However, given the additional level of specificity that such PPIs provide, it is likely that these TF pairs are important for defining tissue-specific expression. Further, these observations are in agreement with conclusion by Ravasi *et al.* that tissue identity is determined not only by specifically expressed TFs, but also strongly relies on tissue-restricted interactions among more generally expressed TFs (Ravasi et al., 2010).

To test this finding more systematically, we evaluated the predictive power of these different levels of specificities: are the combined expressions of individual TFs sufficient to confer tissue-specificity, or do PPIs provide more information? For this, we computed the Pearson correlation coefficients for all gene expression values – excluding TFs – between every pair of 32 tissues. Next we calculated correlation

coefficients (Phi coefficient for binary values) for our classifications of TF specificity, TF presence or absence and PPI presence or absence between each tissue pair. Finally, we applied a linear model to evaluate how well each type of information predicts the correlation in expression levels between tissues.

The results showed that information regarding the presence or absence of PPIs was able to predict tissue-specific expression programmes more accurately than information about individual TFs only (r^2= 0.20, 0.27 and 0.31 for the TFs specfic, TF presence and PPI presence datasets, respectively, Figure 3.6, a, b, c). Therefore, combinatorial TF regulation – in the form of PPIs and their expression – provides very valuable information when looking for differences between tissues. For reference, combining individual and pairwise TF specificity returns the best predictions (r^2= 0.42, Figure 3.6, d).

In summary, our analyses showed that the integration of protein interaction and expression data can greatly advance our understanding of tissue-specific gene regulation.

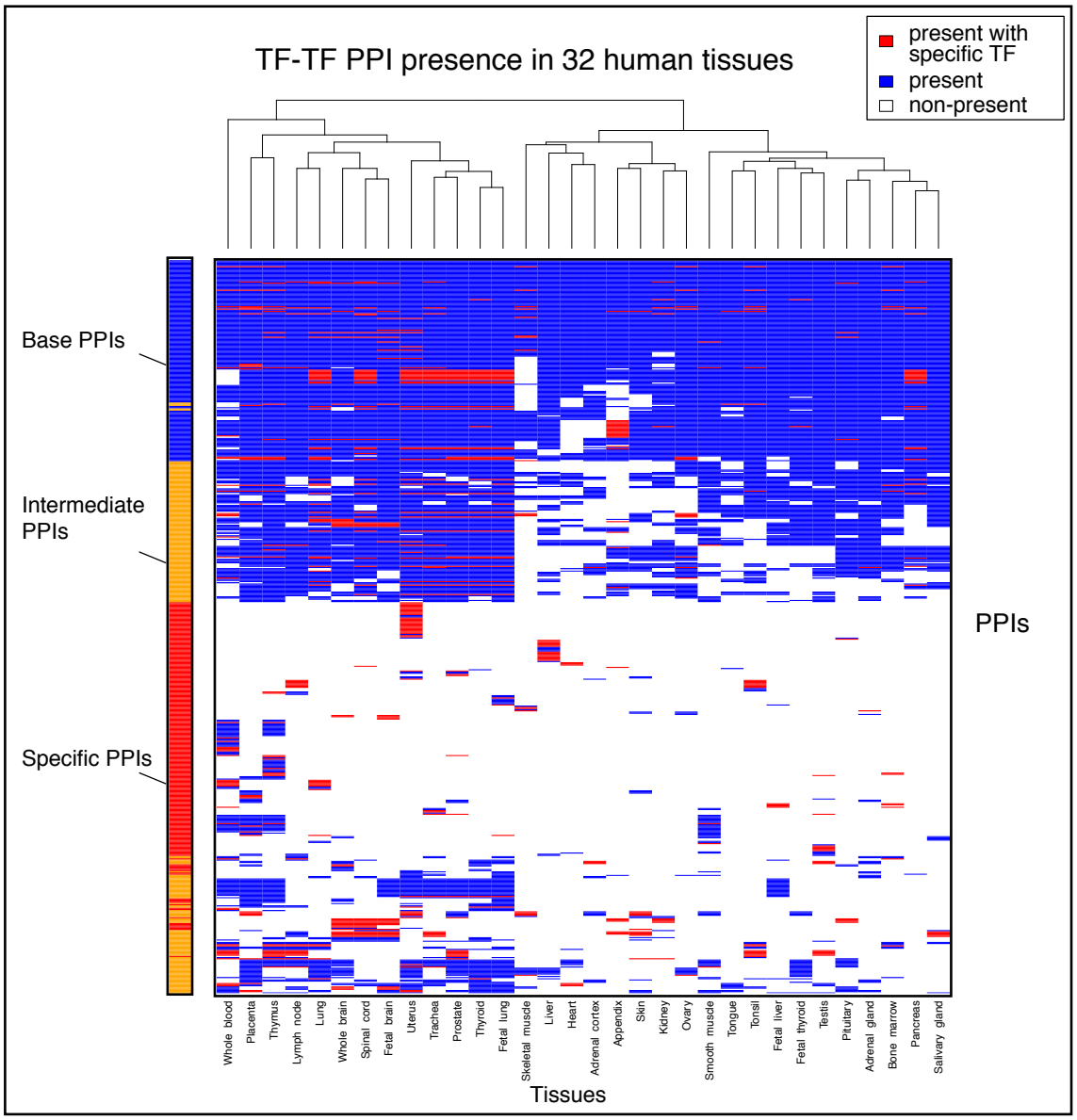


Figure 3.5: TF-TF PPI heatmap. Representation of PPIs (row) present in 32 human tissues (column). PPIs present in a tissue are represented in blue. PPIs in which one of the interacting TFs displays tissue-specific expression are highlighted in red. The bar on the left marks the resulting classification of each PPI based its occurrence in the 32 human tissues (general (blue), intermediate (orange) or specific (red)).

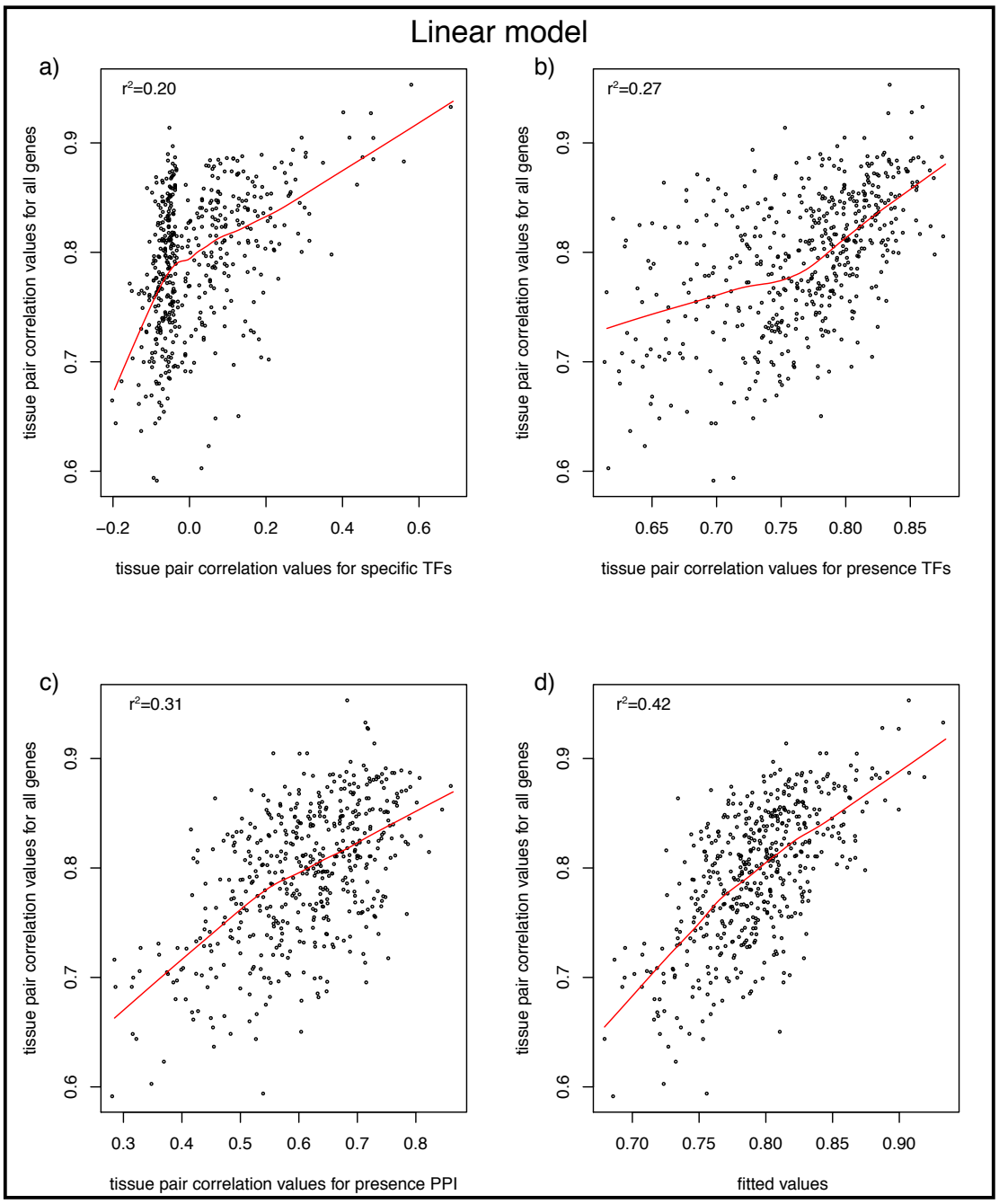


Figure 3.6: Prediction of pairwise tissue expression correlations from TF information. Scatter plots representing pairwise co-expression values between tissues vs pairwise a) specific TFs, b) TF presence or absence and c) PPI presence or absence correlation values. d) result of the linear model when using the TF specificity and PPI presence or absence information to predict the pairwise co-expression values between tissues. The R-squared of each regression is indicated at the top right of the plots.

3.5 Examples of TF networks

Given the predictive power of combinatorial TF-TF pairs, we decided to examine examples of these interactions among different tissues.

First we focused on the lymph node and the tonsil. The two tissues are related, but obviously distinct. They share common functions since they both contain lymphoid tissue: however, the lymph node are more homogenous and are distributed throughout the body. The two tissues display a high degree of correlation of their TF expression (TFs expression Pearson correlation coefficient= 0.93). We observed only a few specific TFs in each tissue—10 and 11 TFs in lymph node and tonsil respectively—of which five are shared. Therefore TF expression alone provides only limited information about the difference between them. When we examined the interactions among TFs however, we identified 245 PPIs that were common between them, but also a large number of PPIs that occur in only one or the other: 70 PPIs are present in the lymph node but not in the tonsil.

Next we compared smooth muscle and the heart TFs (TFs expression Pearson correlation coefficient= 0.93). There are 162 PPIs that are present in both tissues; however we also identified 121 PPIs that are present in the smooth muscle but not heart.

Finally we considered the whole adult and foetal brains. Again, the tissue samples are very correlated at the level of TF expression (TFs expression Pearson correlation coefficient= 0.93), and 10 of the specific TFs are shared between them. When TF-TF interactions were considered, we identified 281 common PPIs; but 35 PPIs were present only in the adult brain and 40 PPIs were present only in the foetal brain. We need to mention that the cell type heterogeneity present in different tissues might affect our analysis. Indeed the different PPIs detected in the different tissues are likely to be linked to different functions performed by the main cell type but as well could reflect the presence of different cell types in the sample.

3.6 Discussion and conclusion

Transcriptional regulation in higher eukaryotes is one of the most basic mechanisms to control cellular responses and define organism complexity. There has been a great deal of research to characterise the proteins that regulate these processes, and to examine how individual regulatory proteins direct different cellular responses. However, even though the combinatorial activity of multiple regulators is considered to be important, relatively few studies have examined this phenomenon on a genomic scale in complex organisms.

In this chapter, we used a recently published dataset of protein-protein interactions and combined it with gene expression data for 32 healthy human tissues. We demonstrated that TFs from the same family tend to interact more than with members of other families. Further, we showed that interacting pairs of TFs were utilised in a hierarchical manner, with some pairs being expressed generally across most tissues and others being expressed specifically. Finally, we showed that incorporating information about TF interactions dramatically increases the predictive power of tissue-dependent gene expression.

A study by Ravasi *et al.* was published while we performed this work; therefore unfortunately most of the results have already been presented elsewhere. The key results that are in common to our studies are the facts that (i) specific TFs tend to be involved in fewer interaction than intermediate and general TFs, (ii) the TF specificity is not the most predominant factor for tissue specificity but this rather relies on tissue-restricted interaction among more generally expressed TFs. In addition, we went into more details in the study of the TFs PPI interactions identifying that most TFs families were enriched for intra-families interactions. Importantly, they identified a small set of 15 highly predictive TF interactions enabling tissues clustering according to their origin. More generally, the fact that many of the observations are reproduced in both studies emphasises the robustness of the

findings. Further, these works represents a fantastic platform to perform many new combinatorial studies on TF-regulation.

There are several important caveats in interpreting these results. First PPI datasets are known to be error-prone, and the data may contain false positives and negatives that might affect the conclusions. Since many of the PPIs we used were detected using low-throughput experiments (reported in the public databases) and could be replicated by different experimental methods (new M2Y set), we believe that the dataset is of good quality. Nevertheless, further improvements to methods for accurately determining PPIs will be important in order to confirm the importance of the TF combinations we observed.

Finally, we emphasise that our analysis is not complete; for example, we did not include co-factors that might interact with DNA-binding TFs. It will be informative to look for these types of regulatory partners that are co-expressed with our TF dataset. This will provide an even more detailed picture of the regulatory complexes that operate within different tissues.

4

MOF-binding and H4K16 acetylation linked to differential gene expression

4.1 Introduction

Maintaining the right level of gene expression is essential for cell survival. To achieve this, cells have evolved complex mechanisms for transcriptional control, and a breakdown in this system can result in diseases or even death. Copy-number variation is one of the most common genetic variants that lead to phenotypic traits; alterations in the numbers of genes can lead to changes in expression. Many diseases including cancer, Williams-Beuren syndrome and Alzheimer's disease have been linked to changes in gene copy numbers (reviewed in Shlien and Malkin, 2009; Zhang et al., 2009). When scaled to a whole-chromosome level, copy-number changes can be responsible for phenotypes such as Down's syndrome, which arises through the presence of an extra copy of all, or part of chromosome 21 (Korbel et al., 2009). Therefore, it is clear that gene dosage - i.e., the number of DNA templates from

which mRNA is produced – plays a crucial role in cell functionality.

Differences in gene-copy numbers also arise naturally as part of the normal life-cycle of an organism, and regulatory mechanisms have evolved to correct for them. One of the best-studied systems is dosage compensation (DC) in which higher eukaryotes compensate for different numbers of sex chromosomes in male and female organisms. Diploid cells have two homologous copies of every chromosome, but the situation is more complex for the sex chromosomes. In mammals and *Drosophila*, females have two X chromosomes whereas males possess only single copies of an X and a Y chromosome. The *C. elegans* worms are males with the X0 sexual chromosome or hermaphrodites with the XX chromosomes. This imbalance is efficiently corrected by DC mechanisms which ensure that genes are expressed at similar levels for X-linked genes across males and females. This means that gene expression must be controlled across an entire chromosome. Therefore, DC represents an excellent model for studying how transcription is coordinated across thousands of genes at a chromosome-wide level.

Traditionally, the molecular study of DC mechanisms has focused on three model systems: worms, flies and mammals. These species have evolved dramatically different methods for DC, suggesting that these mechanisms were acquired relatively late in evolution. Worms correct the dose imbalance between sexual chromosomes by halving the expression of each of the two X chromosomes in hermaphrodites. Mammals on the other hand, randomly inactivate one of the two female X chromosomes during early development. Binding of the non-coding RNA Xist to the entire X chromosome leads to increased histone methylation and partial replacement of histone H2A by the variant macro-H2A. The resulting heterochromatin forms a condensed, transcriptionally inactive structure known as the Barr body (reviewed in Lucchesi et al., 2005).

In contrast, the fly *D. melanogaster* achieves DC by doubling the expression

of genes from the single male X chromosome (Figure 4.1). This is mediated by the activity of the Dosage Compensation Complex (DCC) also known as the Male-Specific Lethal (MSL) complex which recognises and binds to the male X. The importance of the system is highlighted by the fact that deletion of the complex is lethal to male flies.

The DCC is a ribonucleoprotein complex comprising five protein subunits (Male-Specific Lethal subunits 1-3, MSL 1-3; Maleless, MLE; Male Absent On The First, MOF), and two non-coding RNAs (RNA on the X 1 and 2, *roX1-2*). MSLs 1-3 have different binding domains: MSL1 has a PEHE domain, MSL2 contains a RING finger and a cysteine-rich cluster, and MSL3 contains a chromo-related domain and an MRG domain. The non-coding RNAs *roX1* and *2* have different sizes and sequences, but they operate redundantly to target and assist the assembly of the MSL complex on the X chromosome (Meller et al., 2000). The MSL1 and 2 subunits form the core of the DCC, enabling the complex to target a subset of sites on the X chromosome. However the other subunits and the *roX* RNAs are required for full DCC-binding to the X chromosome (reviewed in Gelbart and Kuroda, 2009; Hallacli and Akhtar, 2009). In females, MSL2 translation is prevented by the Sex-Lethal protein (SXL), so blocking the formation of the MSL complex.

MLE and MOF are enzymes: MLE is an RNA/DNA helicase containing a double-stranded RNA-binding domain, and MOF is a MYST histone acetyltransferase (HAT). MOF specifically acetylates lysine 16 on the histone H4 molecule (Smith et al., 2000; Akhtar and Becker, 2000). By targeting MOF-binding, the DCC promotes H4K16 acetylation on the X chromosome (Hilfiker et al., 1997; Akhtar and Becker, 2000; Smith et al., 2000, 2001).

Histone acetylation is almost invariably associated with increased transcription in many different organisms. In general, acetylation is thought to weaken histone-DNA interactions by neutralising the positive charge on lysine residues. More specifically,

H4K16ac has been shown to contribute to decompaction of the 30nm chromatin fibre (Shogren-Knaak et al., 2006; Robinson et al., 2008). Removal of the H1 linker histone leads to further decompaction. The more open chromatin structure provides a favourable environment for gene expression. In budding yeast, over 80% of H4 subunits are acetylated on K16, and the mark is shown to maintain or promote transcription in vivo (Suka et al., 2002; Kimura et al., 2002). Further, increased levels of H4K16ac at promoter regions have been shown to de-repress transcription in yeast (Akhtar and Becker, 2000). Therefore, this chromatin mark is a very strong indicator for transcriptional activity.

Following identification of the MSL proteins, studies have focused on characterising the DC mechanism. An important step towards this was the identification of the binding locations of the MSL complex. Early studies utilised immunostaining of polytene chromosomes from salivary glands (Kelley et al., 1999; Bone et al., 1994); while the method provides an excellent global view of binding, it has relatively low resolution and therefore insufficient for determining whether particular genes are bound or acetylated. ChIP-chip experiments of the MSL complex and the histone mark have provided genome-wide maps of binding at very high resolution. These experiments were usually performed in male and female cell lines in parallel, and lead to the following findings: (i) the MSL complex preferentially binds to the male X chromosome; (ii) the MSL1 subunit targets about 50% of all X-linked genes; (iii) MSL1 and MSL3 preferentially localise to the 3'-end of X-linked genes; (iv) but surprisingly the enzyme MOF also binds to the 5'-end of genes across all chromosomes, giving rise to a bimodal binding pattern on the X chromosome, but a skewed 5'-end distribution on autosomes; (v) patterns of H4K16 acetylation correlate strongly with MOF binding; (vi) genes bound by the MSL complex are generally transcriptionally active—though not all expressed genes are bound by the complex (Smith et al., 2001; Legube et al., 2006; Alekseyenko et al., 2006; Gilfillan et al., 2006; Kind

et al., 2008; Gelbart et al., 2009). Recent studies by the Akhtar group have shown that another protein complex, Non-Specific Lethal (NSL) is responsible for directing MOF-binding to the 5'-end of autosomal and X chromosomal genes (Mendjan et al., 2006; Raja et al., 2010).

These binding studies led to the proposal and refinement of a two-step model for MSL function (initially proposed by Kelley et al., 1999). First the MSL complex recognises and binds to high-affinity sites on the X chromosome, also known as chromatin entry sites. Sequence analysis has revealed that these sites consist of near-perfect repeats of eleven GA dinucleotides (Alekseyenko et al., 2008; Straub et al., 2008). Binding by the complex then spreads to neighbouring regions until most of the X chromosome is occupied (reviewed in Gelbart and Kuroda, 2009). In conjunction with this spreading, histones are covalently modified with marks that are indicative of transcriptional activity (Larschan et al., 2007; Bell et al., 2008; Sural et al., 2008).

Although these findings have provided good insights into the mechanisms for DC, there are still some unanswered questions. An important topic is the precise regulatory function of MOF: what are the consequences of MOF-binding and the effects of H4K16 acetylation on gene expression?

In chapters 4 and 5, we present a study performed in collaboration with the Akhtar laboratory. To assess MOF function, we performed experiments in live tissues (salivary gland of 3^{rd} instar lavae) from male and female flies rather than cell lines, as is usually the case. Salivary gland euchromatin undergoes endoreplication, up to 1024 copies to form the polytene chromosomes. Notably for our analysis, inside each cell the X to autosome ratio is kept after these replications — this is important for the cell to function normally and has also been verified by qPCR. ChIP-seq experiments measured MOF-binding and the pattern of H4K16 acetylation. Microarray experiments measured changes in gene expression. Two MOF mutant strains were

used, *mof1* (Hilfiker et al., 1997) and *mof2* (Gu et al., 1998); of these, the second produces a truncated protein that completely abolishes MOF's enzymatic activity. This provides much greater sensitivity for characterising MOF's function compared with RNAi-mediated knock-downs, which leads only to incomplete abolition of MOF activity.

In analysing these data, we aimed to assess the role of MOF with respect to H4K16ac and gene expression. In this chapter, we first describe the pattern of MOF-binding and H4K16 acetylation in male and female wt tissues. We then examined how acetylation changes in the absence of a functional MOF protein. Finally we explained the changes in gene expression in mutant flies.

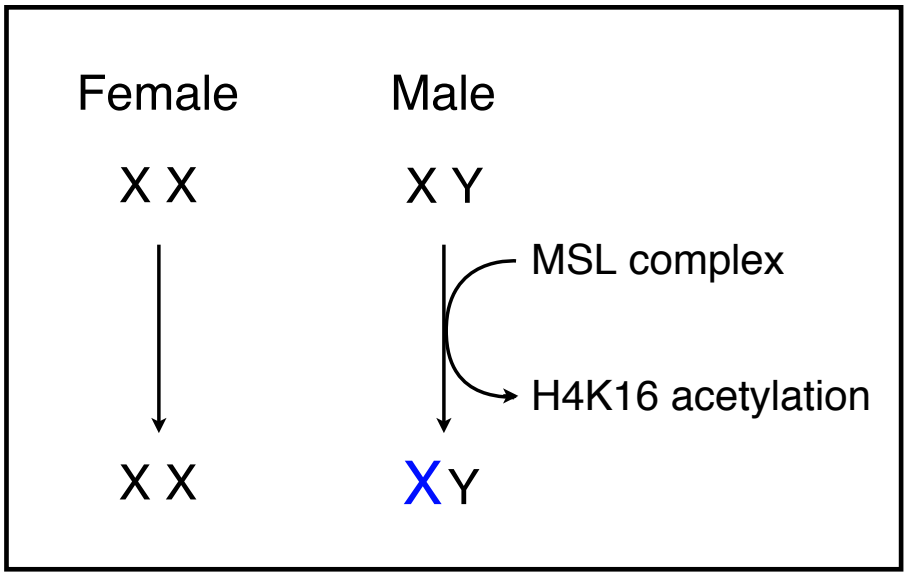


Figure 4.1: Schematic figure describing dosage compensation in *D. melanogaster*. X and Y denote the two sexual chromosomes. The approximate two-fold increase in gene expression from the single male X chromosome is mediated by the MSL complex which acetylates H4K16. (Adapted from Mendjan and Akhtar, 2007).

4.2 Material and Methods

4.2.1 Biological materials

All the analyses presented in chapters 4 and 5 are based on unpublished ChIP-seq and microarray data. All the experiments presented here were performed by Thomas Conrad, a PhD student in the Akhtar Group; all the computational analysis was performed by me.

Biological samples originated from the salivary glands of third instar male and female larvae in wt, *mof1* and *mof2* mutant backgrounds. *Mof1* mutants carry a point mutation in the MOF gene that affects the catalytic domain (G691E), so reducing acetylation activity by 90% (Akhtar and Becker, 2001). Mutant male flies die before adulthood; 80% of female mutants reach adulthood, but their average lifespan is reduced to 29 days compared with 37 days in wt. *Mof2* mutants carry a premature stop codon (Q371X) which generates a truncated, non-functional protein. Mutant male flies do not reach adulthood; 54% of females do so, but with an average lifespan of 7.8 days. Figure 4.2 shows a schematic of the MOF gene depicting its protein-coding domains as well as showing the locations of the point mutations.

4.2.2 Microarray dataset

To assess gene expression, we hybridised mRNA extracted from male and female wt, *mof1* and *mof2* salivary glands to Affymetrix *Drosophila* 2 GeneChips. Experiments were performed as three biological replicates except for the male wt for which two sets of triplicates were produced. RNA samples were processed according to standard manufacturer instructions.

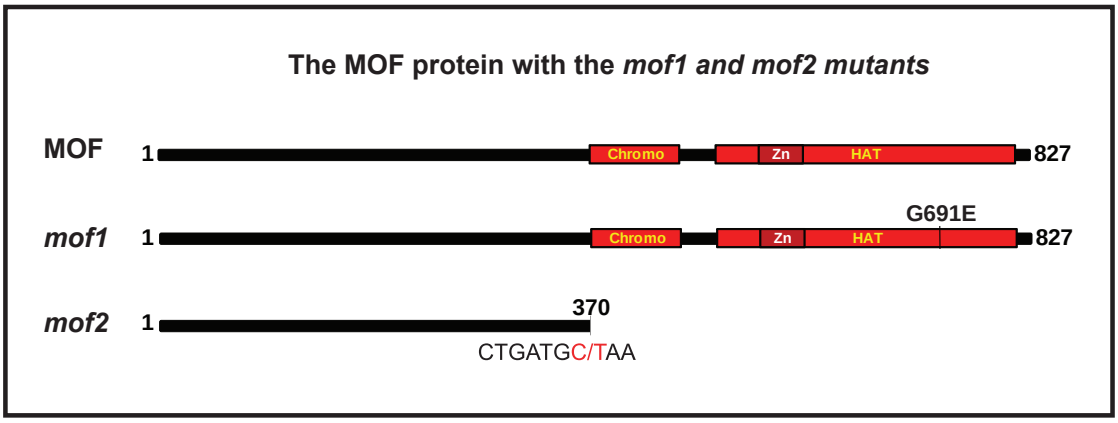


Figure 4.2: Schematic representation of the MOF protein and mutants used in this study. The upper panel shows the protein domain structure of MOF predicted by Interpro: Chromo, chromo-barrel domain; Zn, zinc finger domain; HAT, Acyl_CoA_acyltransferase. The middle panel shows the *mof1* mutant with a point mutation at position 691 within the catalytic site (G691E). The lower panel shows the truncated MOF protein in the *mof2* mutant resulting from a point mutation at position 371 introducing a stop codon (Q371X). The underlying nucleotide sequence change is given below with the mutated base in red. Bold numbers at the beginning and end of each protein indicate the reference positions for first and last amino acids.

4.2.3 ChIP-seq dataset

To identify MOF-binding and H4K16 acetylation sites, we performed ChIP-seq experiments in male and female wt, and *mof2* flies. Table B.1 in the appendix summarises all the ChIP-seq experiments analysed, and provides the numbers of sequenced and mapped reads for each sample. Owing to rapid improvements in the sequencing technology, the initial 36nt reads were lengthened to 72nt reads in later samples. Anti-MOF antibodies came from stocks in the Akhtar laboratory, while anti-H4 and H4K16ac antibodies were purchased from Abcam and Santa Cruz Biotechnology, respectively.

ChIP-seq experiments were not performed for female *mof2* mutants, as it was not possible to generate sufficient numbers of homozygous flies to obtain enough material. The *mof* gene is located on the X chromosome, and male *mof2* mutants—which die during development—are needed to produce offspring.

Replicate experiments were performed only for Pol II-binding presented in chapter 5 because of the high sequencing costs. However, data for those experiments

showed good reproducibility indicating high data quality (ρ=0.74 for male, ρ=0.68 for female, section 5.2.4.1 and Figure 5.1 in chapter 5).

4.2.4 Microarray data analysis

4.2.4.1 Quality check and normalisation

Microarray data analysis was performed using the R statistical software (R Development Core Team, 2008; Ihaka and Gentleman, 1996) and packages from the Bioconductor project (Gentleman et al., 2004). We evaluated the quality of the 18 GeneChip hybridisations using the ArrayQualityMetrics package (Kauffmann et al., 2009), which assesses the quality of individual arrays and identifies apparent outlier arrays. One male wt array was discarded because it did not pass the quality threshold. Next, we normalised data from the 17 remaining arrays using GCRMA (Gautier et al., 2004; Wu et al., 2004). The three-step algorithm consists of: (i) background correction of each array with an affinity model that takes into account the probe sequence; (ii) across-array quantile normalisation which removes non-biological systematic variation using a ranking procedure; and (iii) median polish summarisation which produces robust estimates of probeset expression by correcting for probe and array-specific effects using a linear model fit to the data. This procedure outputs normalised log2 expression values for each probeset on the array.

Using annotations from Ensembl Genomes, we mapped 13,905 probe sets to 12,715 genes on the fly genome (10,663 autosomal and 2,052 X chromosomal genes). 5,047 probesets were removed from the analysis as they mapped to more than one gene.

4.2.4.2 Differential expression analysis

Differential expression analysis was performed using the Limma package within Bioconductor (Smyth, 2005). The package implements an empirical Bayes approach to calculate a global variance estimator based on all gene variances. It uses a moderated t-test as the test statistic, in which the single-gene estimated variance is replaced by a weighted average of the global and single gene variances. Under certain distributional assumptions, this statistic follows a t-distribution under the null hypothesis with the degrees of freedom depending on the dataset. We applied this method to our data using an adjusted p-value threshold of 0.05. We identified significantly differentially expressed genes (up- and down-regulated genes) in the mutant samples compared with wt samples in both male and female microarray data sets.

4.2.5 ChIP-seq data analysis

4.2.5.1 Data processing

ChIP samples were sequenced using Illumina GAIIx machines at the EMBL GeneCore facility. Resulting reads were mapped to the *D. melanogaster* genome (`dmel_r5.11_FB2008_08`) using the Bowtie software (Langmead et al., 2009) with the following parameters: -n -k 1 –solexa1.3-quals –best. With these settings, the software maps reads only to unique locations in the genome, choosing the best possible hit. In order to return the best hit, the algorithm accounts for read-quality during alignment. Furthermore, reads containing ambiguous nucleotide assignments (N) were removed. Finally, we considered only reads mapping to chromosomes 2L, 2R, 3L, 3R, 4 and X (Table B.1).

We divided the *D. melanogaster* genome into non-overlapping 25 bp bins, and we counted the number of reads mapped to each bin for each ChIP-seq sample. We inputed these read counts into the DESeq Bioconductor package (Anders and

Huber, 2010); counts were then normalised between IP and input samples by applying a scaling factor that accounts for differences in the total numbers of reads per sequencing run. Any bins with read counts of zero—in either the IP, the input or both samples—were discarded from further analysis. For each bin, DESeq outputed a log2 fold-change (log2FC) value between normalised read counts in the IP and control samples. We used input DNA as the controls for MOF- and Pol II-binding and histone H4-binding as the control for the H4K16 acetylation marks. As DNA fragment sizes after sonication was $\approx$ 200 bp, we smoothed log2FC values using a 400 bp sliding window approach (Kind et al., 2008). The final log2FC values represented the signal from the IP samples relative to controls, with positive values corresponding to enrichments in the IP sample and negative values corresponding to enrichments in control sample.

We calculated log2FC thresholds to indicate significant binding or acetylation compared with the control. Since negative log2FC values (i.e. enrichment in control signal) correspond to experimental noise, we fitted a symmetric null-distribution to the density distribution for log2FC values below the mode. We then applied an FDR-adjusted p-value cut-off of 0.05 to identify 25 bp bins containing significant binding or acetylation (Schwartz et al., 2006). Figure 4.3 shows density distributions of log2FC values for MOF-binding in male and female wt flies, and the significance thresholds for them. Similar distributions for the H4K16ac IP experiments are presented in Appendix B.1.

4.2.5.2 Detection of MOF-bound and H4K16-acetylated genes

We used the pattern of MOF-binding to classify genes into three categories: (i) those with no binding (not); (ii) those that are partially bound - typically in the promoter region (prom); (iii) those that are fully bound along their whole length (full).

The gene body was defined as all exonic sequences between +500 bp downstream

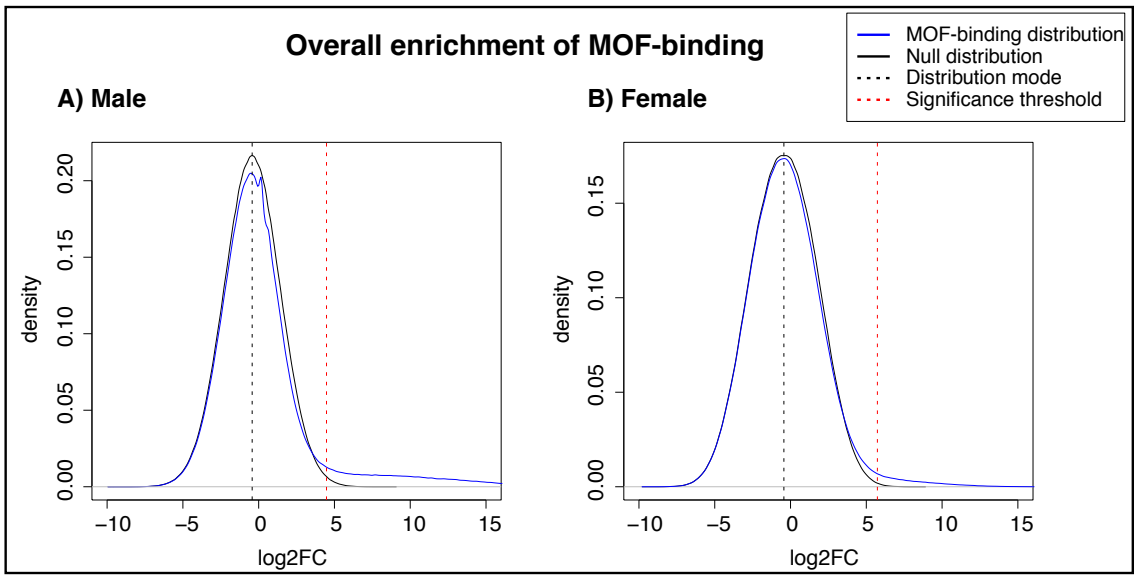


Figure 4.3: The overall enrichment of MOF-binding in male and female. The density plots show the distributions of log2FC (IP/input) values (blue lines), the fitted null distributions (black lines) and significance thresholds (red dashed lines) for MOF-binding in wt males (A) and females (B). The modes of the empirical distributions are indicated by black dashed lines.

of the transcription start site (TSS) to the 3'-end, as annotated in the Ensembl database. The promoter was defined as the region between -200 bp and the TSS: the region was selected by identifying the mode for average MOF-binding at the 5'-end of all genes—which falls about -100 bp upstream of the TSS in both male and female samples—and then providing a 100 bp window either side (Figure 4.5). We excluded any genes that are <600 bp (1,205 loci) in length, as they were too short for this analysis.

Figure 4.4 shows the distribution of the percentage MOF-binding in gene bodies in the male wt sample. There is a clear bi-modal distribution dividing partially bound genes from fully bound ones. We used a threshold of 70% (relative to the gene body length) to differentiate between the two types of binding. Next, we classified partially bound genes as promoter bound, if they contained at least one bin with significant binding in the promoter region defined above.

We used a similar classification to identify patterns of H4K16 acetylation. In this case, the promoter was defined as the region between the TSS and +500 bp

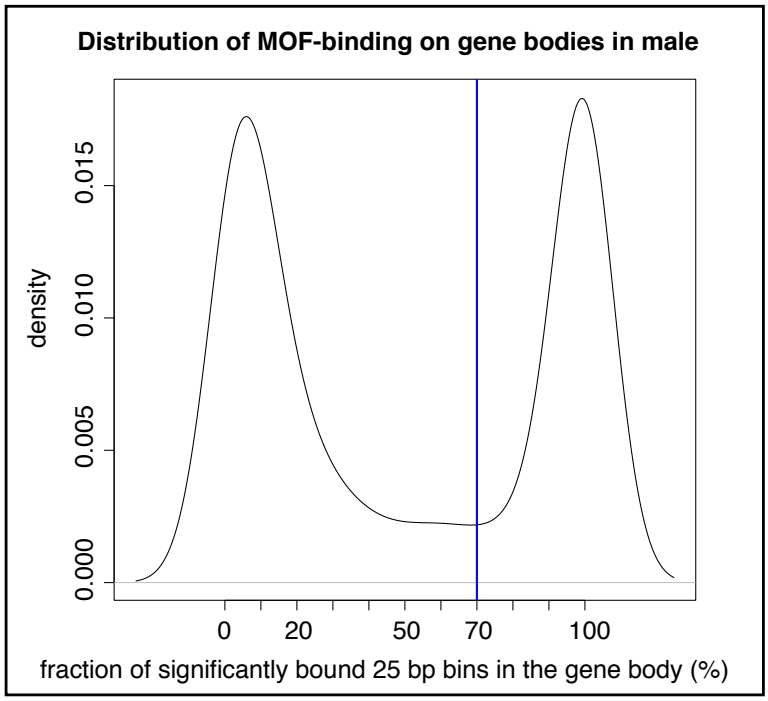


Figure 4.4: The distribution of MOF-binding on genes bodies in male flies. The density plot shows the distribution of genes with a given percentage of gene bodies bound by MOF. We set a 70% threshold to distinguish between fully bound and partially bound genes. A similar plot for female wt flies is presented in Appendix B.2.

downstream, in order to accommodate shift in acetylation patterns towards the 3'-end of genes compared with MOF-binding (Figure 4.5).

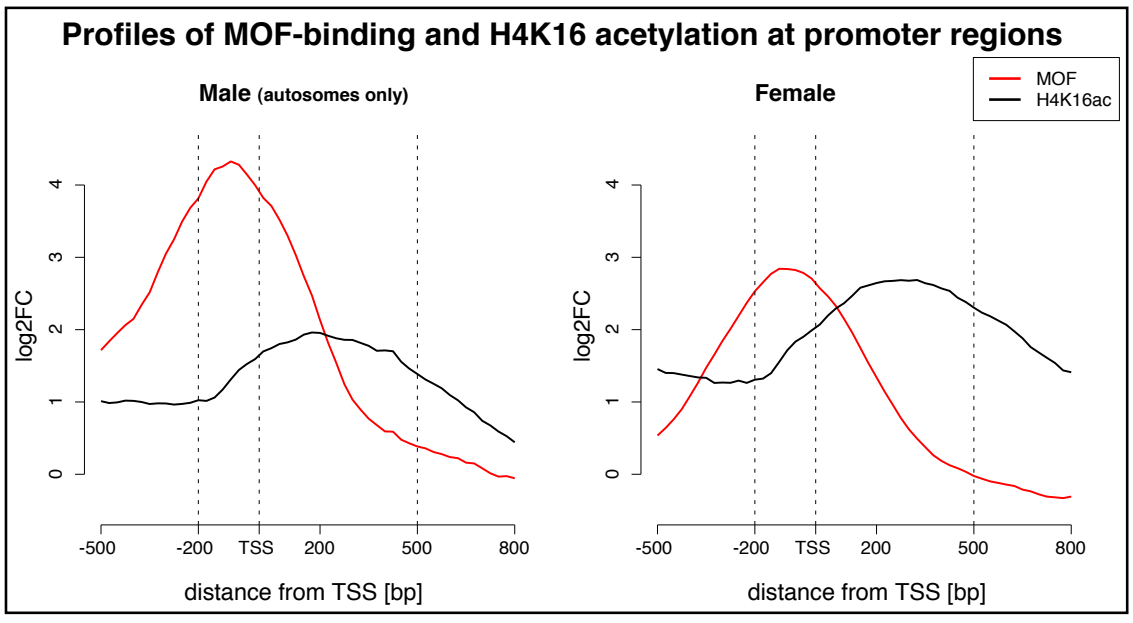


Figure 4.5: Profiles of MOF-binding and H4K16 acetylation at promoter regions. The density plots show that average log2FC signal for MOF-binding (red) and H4K16ac (black) around the transcriptional start site in wt male (left) and female (right) samples. For males, we included only autosomal genes. The dashed lines delimit the promoter regions defined for MOF-binding and H4K16 acetylation.

4.3 Characterisation of MOF-binding and H4K16 acetylation

4.3.1 MOF-binding and H4K16 acetylation display different profiles between the male X chromosome and autosomes

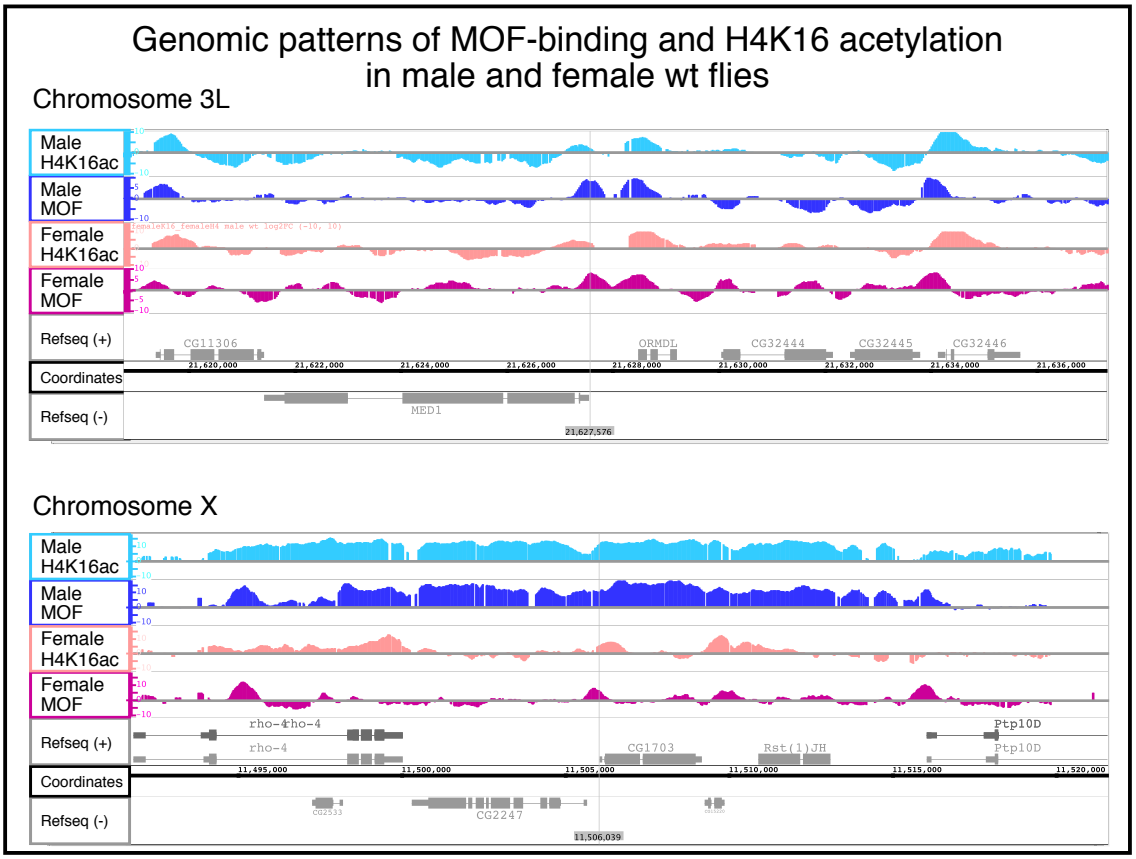


Figure 4.6: Genomic patterns of MOF-binding and H4K16 acetylation. Screenshots of the Integrated Genome Browser (Nicol et al., 2009) show the patterns of MOF-binding (dark colour) and H4K16 acetylation (light colour) in male (blue colour shades) and female (red colour shades) salivary glands along a 16kb section of chromosome 3L (top) and a 25 kb section of the X chromosome (bottom). Genomic coordinates and the locations of annotated Refseq genes on the forward (+) and reverse (-) strands are indicated below.

Figure 4.6 displays the patterns of MOF-binding and H4K16 acetylation on sections of chromosomes 3L and X in male and female wt samples. Visual inspection of

these profiles confirms that the two signals overlap significantly. It is also clear that there are distinct patterns of MOF-binding and acetylation on the X chromosome compared with autosomes in males.

To confirm that these observations apply generally, we computed composite binding and acetylation profiles by averaging log2FC values across all genes. These distributions highlight two distinct patterns of binding: (i) on the male X chromosome, MOF binds to the entire gene body, with peaks of increased binding at the promoter and the 3'-end; (ii) on male autosomes and in females, MOF binds only at gene promoters. The pattern of H4K16 acetylation follows that of MOF-binding quite closely; interestingly there is a shift in signal towards the 3'-end, reflecting the fact that MOF acetylates histones that are downstream of its binding site (Figure 4.7).

The context-dependent behaviour of MOF – i.e., the distinct binding patterns depending on the chromosome – is in agreement with previous investigations of MOF function (Kind et al., 2008). The agreement of our observations with previous publications indicates that our data are reliable.

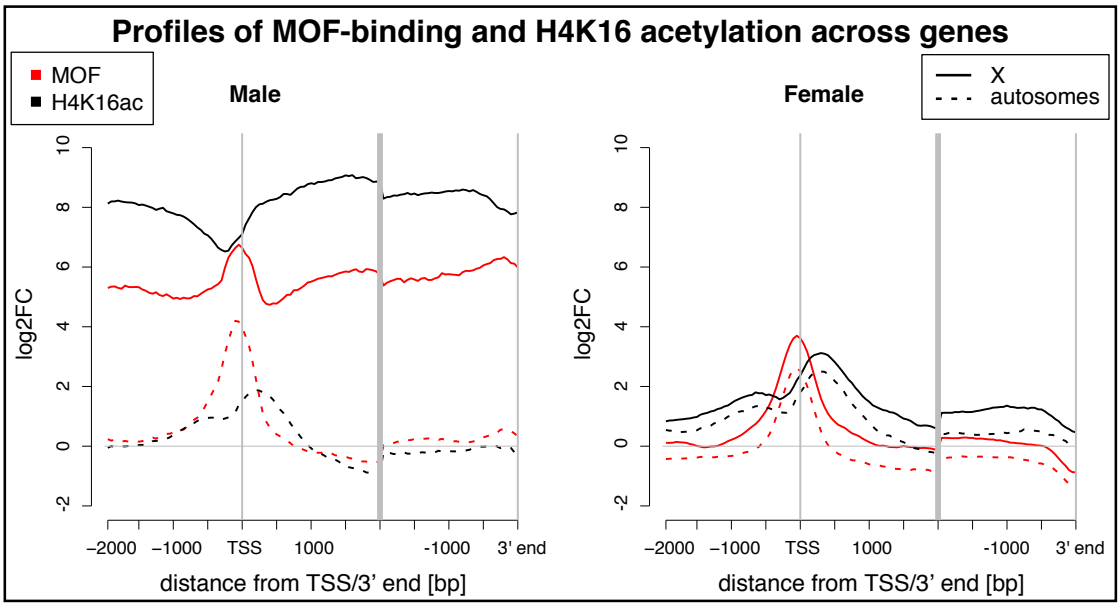


Figure 4.7: Profiles of MOF-binding and H4K16 acetylation along genes. The plots show the average log2FC signal for MOF-binding (red) and H4K16ac (black) along genes in wt males (left) and females (right). Log2FC signals are shown for the region surrounding the TSS (-2000bp to +2000bp) and for the 3'-end of the gene (3'-end to -2000bp upstream of termination site). Average log2FC values for X-linked and autosomal genes are represented by solid and dashed lines, respectively.

4.3.2 Identification of MOF-bound and H4K16-acetylated genes in wt salivary glands

Having characterised the global patterns of MOF-binding and H4K16 acetylation, we next classified sets of bound and acetylated genes. We used the criteria described above (section 4.2.5.2) to define: (i) fully MOF-bound or acetylated genes; (ii) promoter-bound or acetylated genes and (iii) unbound or un-acetylated genes. Table 4.1 summarises the numbers of genes from wt male and female flies in each category.

In males, 898 (40%) X-linked genes were fully bound by MOF, and 1,704 (76%) genes were fully acetylated. In contrast in females, only one X-linked gene was fully bound and 158 (7%) were fully acetylated, demonstrating that extensive MOF-binding and acetylation on the X chromosome are specific to males. These genes

A)

Male MOF	X-linked	(%)	Autosomal	(%)	Total	(%)
Total bound	1,285	57.65	4,830	41.19	6,115	43.82
- gene body	898	40.29	50	0.43	948	6.79
- promoter	387	17.36	4,780	40.77	5,167	37.03
Not bound	944	42.35	6,895	58.81	7,839	56.18
Total	2,229		11,725		13,954	

B)

Female MOF	X-linked	(%)	Autosomal	(%)	Total	(%)
Total bound	903	40.51	3,862	32.94	4,765	34.15
- gene body	1	0.04	8	0.07	9	0.06
- promoter	902	40.47	3,854	32.87	4,756	34.08
Not bound	1,326	59.49	7,863	67.06	9,189	65.85
Total	2,229		11,725		13,954	

C)

Male H4K16ac	X-linked	(%)	Autosomal	(%)	Total	(%)
Total H4K16ac	1,989	89.23	4,690	40	6,679	47.86
- gene body	1,704	76.45	321	2.74	2,025	14.51
- promoter	285	12.79	4,369	37.26	4,654	33.35
No H4K16ac	240	10.77	7,035	60	7,275	52.14
Total	2,229		11,725		13,954	

D)

Female H4K16ac	X-linked	(%)	Autosomal	(%)	Total	(%)
Total H4K16ac	1,064	47.73	5,023	42.84	6,087	43.62
- gene body	158	7.09	682	5.82	840	6.02
- promoter	906	40.65	4,341	37.02	5,247	37.6
No H4K16ac	1,165	52.27	6,702	57.16	7,867	56.38
Total	2,229		11,725		13,954	

Table 4.1: Numbers and percentages of MOF-bound and H4K16-acetylated genes. Numbers and percentages of MOF-bound (A, B) and H4K16-acetylated (C, D) genes on the X chromosome and the autosomes in wt males (A, C) and females (B, D). The important numbers for the male-female comparison are printed in red.

are most likely to be dosage compensated. On autosomes, in males about 40% of genes are bound and acetylated at the promoter (4,830 and 4,690 genes respectively). Roughly similar numbers of genes are affected in females (3,862 (33%) and 5,023 (43%) genes bound and acetylated, respectively).

On comparison with ChIP-chip data published by Kind et al. (2008), we detected

a significant overlap in the sets of MOF-bound and acetylated genes: 91% (male, 2,675 genes) and 76% (female, 2,232 genes) of genes reported as MOF-bound by Kind were present in our dataset. In addition, we found further 3,440 and 2,533 MOF-bound genes in male and female respectively, reflecting the fact that the ChIP-seq techniques used here are more sensitive that ChIP-chip. It is worth noting that the Kind study used karyotypically unstable cell lines; therefore the large overlap despite differences in the biological samples highlights the robustness of these observations.

4.3.3 MOF-bound genes display higher levels of H4K16 acetylation

To characterise the overlap between patterns of MOF-binding and H4K16 acetylation, we compared the lists of classified genes above (section 4.3.2). On autosomes, we found that 83% (males) and 87% (females) of MOF-bound genes were also acetylated (Figure 4.8, A). Therefore there was good correlation between MOF-binding in promoters, and the pattern of acetylation.

Similarly on the X chromosome, a large proportion of MOF-bound genes in male (1,254 genes, 98%) and female (761 genes, 84%) samples were acetylated (Figure 4.8, B). However in males, we also detected 735 genes that were acetylated despite lacking MOF, suggesting spreading of H4K16 acetylation by MOF to neighbouring genes. This is in agreement with the model in which acetylation spreads across genes on the male X chromosome (Kelley et al., 1999; Kind et al., 2008; Gelbart et al., 2009). We did not observe the same phenomenon on the female X chromosome.

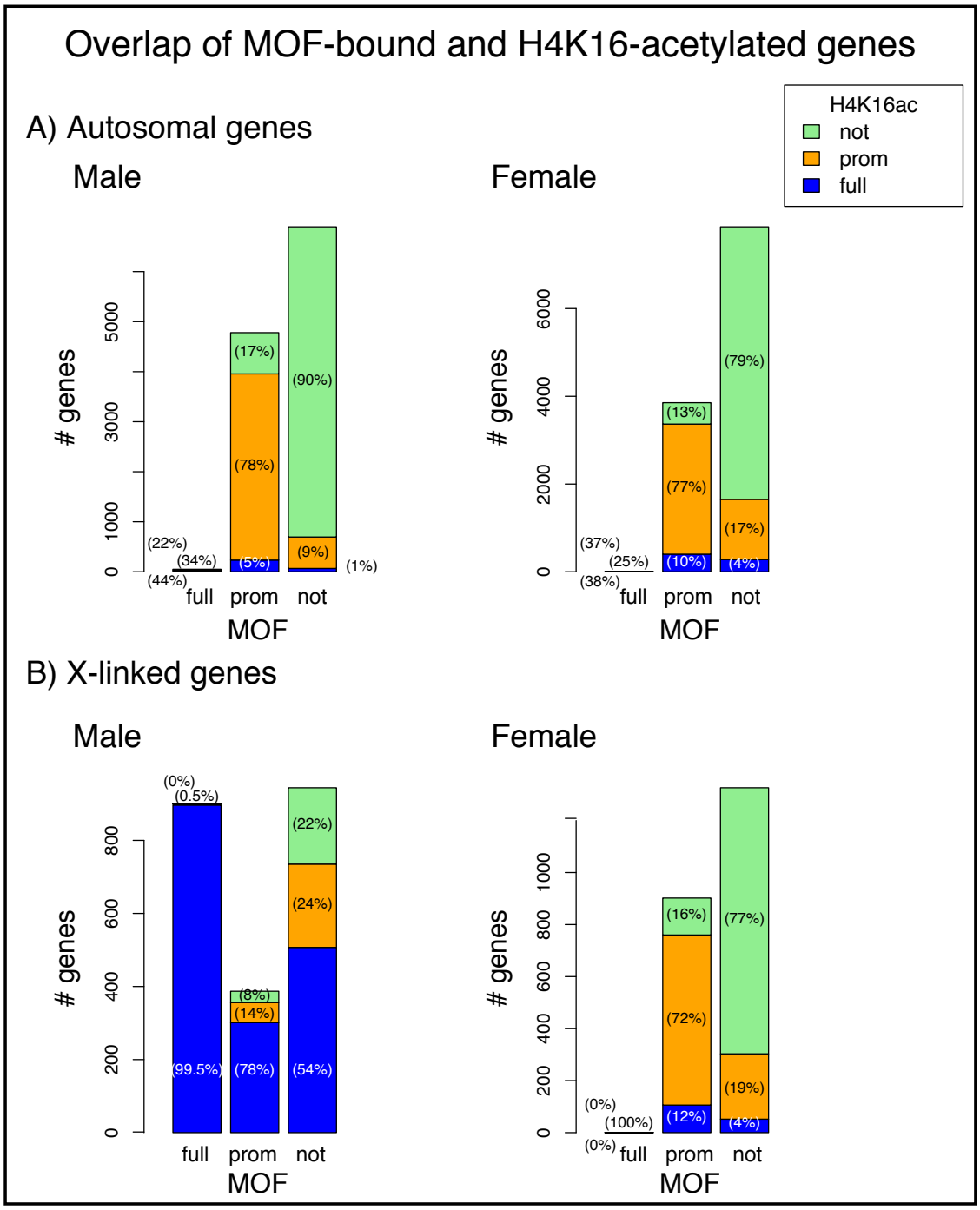


Figure 4.8: Overlap of MOF-bound and H4K16-acetylated genes. Autosomal (A) and X-linked (B) genes were divided into three categories showing full (full), promoter-only (prom) or no (not) binding by MOF. For each category, the stacked barplot specifies the number of genes showing full (full), promoter-only (prom) or no (not) H4K16 acetylation. The numbers within each block indicate the percentage of genes within this block relative to all genes in the category.

4.3.4 Male and female autosomes show similar MOF-binding and H4K16 acetylation patterns

Next we compared the pattern of MOF-binding and acetylation between males and females. This comparison allowed us to examine the effects of DC on the single male X chromosome. We used two approaches for this: (i) evaluating the correlation in significant log2FC values for MOF-binding and H4K16ac; and (ii) comparing the sets of MOF-bound and acetylated genes.

Since DC does not occur on autosomes, we expected to detect similar levels of binding and acetylation between males and females. Since MOF preferentially binds on expressed genes (Legube et al., 2006; Alekseyenko et al., 2006), the only exceptions should be genes with sex-specific functions and few such genes were expected to be expressed in salivary glands. As shown in Figure 4.9 (A & B), we observed high levels of correlation for both binding and acetylation in autosomal promoters (ρ=0.77 and ρ=0.86 respectively). Additionally, we found large overlaps between the groups of classified genes: 98% (3,801 genes) of MOF-bound genes in females were also bound in males, and 87% (4,394 genes) of acetylated genes in females were also acetylated in males.

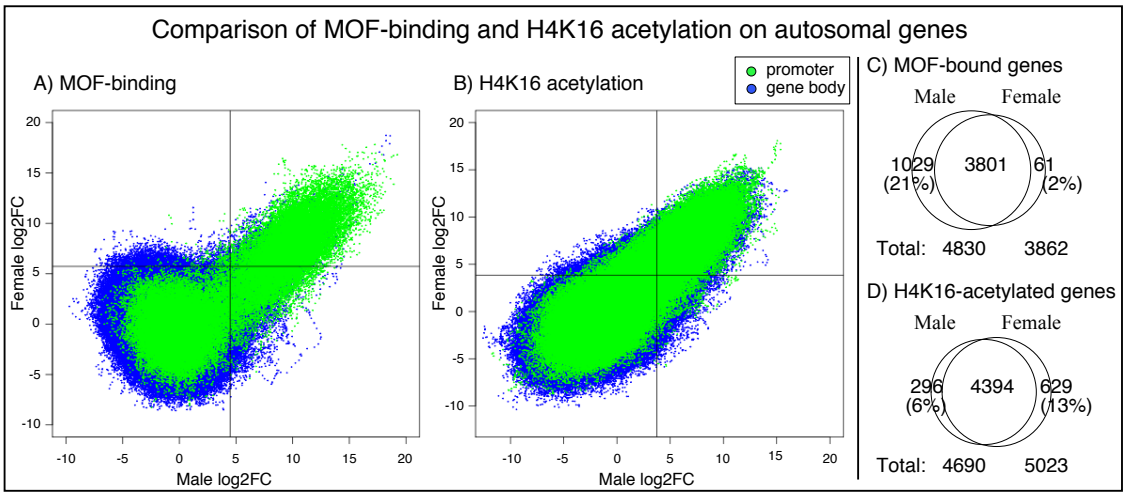


Figure 4.9: Comparison of MOF-binding and H4K16 acetylation on autosomal genes in males and females. (A, B) Scatterplots comparing log2FC values for MOF-binding (A) and H4K16 acetylation (B) in autosomal genes in males (x-axis) and females (y-axis). Blue and green dots represent log2FC values from gene bodies and promoter regions, respectively. Black vertical and horizontal lines represent the MOF-binding (in A) and H4K16 acetylation (in B) significant thresholds for male and female, respectively. (C, D) Overlap in MOF-bound (C) or H4K16-acetylated (D) autosomal genes between males and females (total numbers of genes given below).

4.3.5 MOF and H4K16 acetylation show a distinct binding pattern in the male X chromosome

Since the X chromosome is subject to DC, we expected to observe large differences in patterns of MOF-binding and H4K16 acetylation. Figure 4.10 (A & B) shows the correlation of MOF-binding and H4K16ac between males and females.

Given the dual-binding pattern of MOF at promoters and gene bodies on the male X chromosome, we compared the amount of binding for these gene regions separately. At promoters where MOF binds both male and female genes, we found a high level of correlation (ρ=0.72). However, this correlation was lost when binding was compared for gene bodies (ρ=0.05), highlighting a difference in MOF function between males and females. Similarly, for H4K16ac, we found a correlation for gene promoters (ρ=0.54), but no correlation along the body of genes (ρ=0.27).

Then, we evaluated the overlap between the bound and acetylated genes in males

and females (Figure 4.10, C & D). Using the set of all bound genes, (i.e., fully bound and promoter bound), we found that almost all female MOF-bound genes (99%) were also bound in males. A similar result was observed for H4K16 acetylation. However, a significant proportion of genes were MOF-bound (31%, 395 genes) or acetylated (47%, 936 genes) only in males.

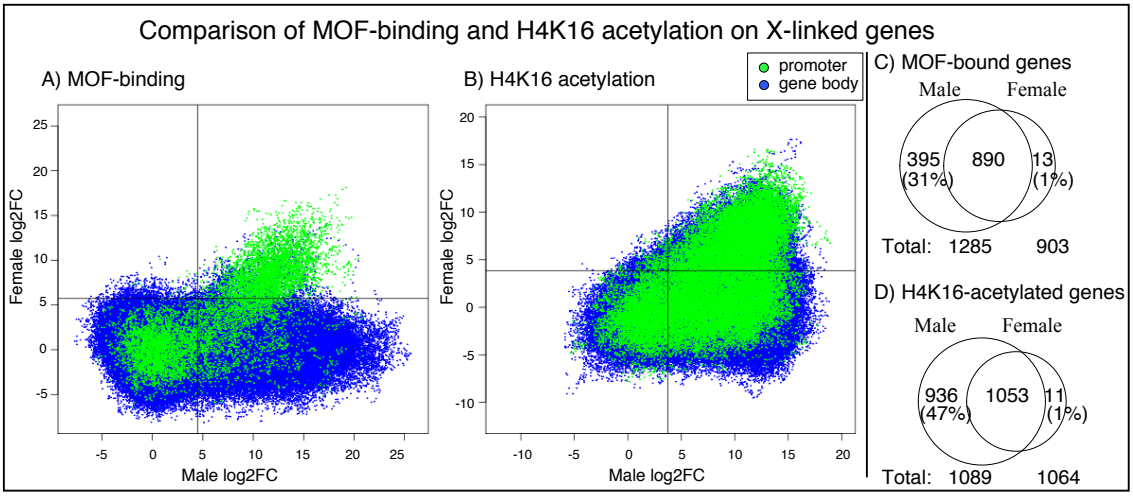


Figure 4.10: Comparison of MOF-binding and H4K16 acetylation on X-linked genes in males and females. (A, B) Scatterplots comparing log2FC values for MOF-binding (A) and H4K16 acetylation (B) on X-linked genes in males (x-axis) and females (y-axis). Blue and green dots represent values from gene bodies and promoter regions, respectively. Black vertical and horizontal lines represent the MOF-binding (in A) and H4K16 acetylation (in B) significant thresholds for male and female, respectively. (C, D) Overlap of MOF-bound (C) or H4K16-acetylated (D) X-linked genes between males and females (total numbers of genes given below).

4.4 The role of MOF in regulating gene expression

4.4.1 Large down-regulation of X-linked genes in male *mof2* mutants

Next we evaluated the effects of MOF-binding on gene expression levels in male and female flies. For this, we measured expression levels in wt males and females, and

compared them with expression levels in male and female *mof2* mutant flies. By comparing wt and mutant expression profiles, we are able to assess the effect that the loss of MOF activity has on gene expression.

For quality control, we examined the amount of differential expression between wt male and female samples. Given that the salivary gland is unrelated to sex-specific functions, we expected most genes to be similarly expressed. This was indeed the case, and we identified only 175 (1.5%) differentially expressed genes between the samples (Table B.2, a) in appendix B).

We then examined the amount of differential expression between wt and mutant samples. Given that MOF functions differently on autosomes and the X chromosome, we analysed them separately. Figure 4.11 displays the percentage of differentially expressed genes in male and female mutants compared with the wt samples (see also Table B.2, b) & c) in the appendix B).

We detected a pronounced down-regulation of X-linked genes in the male *mof2* mutant (403 genes), which most likely reflects the loss of DC. In addition, we detected a large amount of up-regulation among autosomal genes, which we will discuss later.

These observations were validated by qRT-PCR on a subset of 15 genes using genomic DNA as a reference (Figure 4.12). In every case, the qRT-PCR experiments recovered similar trends as the microarray data.

These results demonstrate that loss of MOF function leads to a dramatic down-regulation of X-linked genes in males. In contrast, genes on autosomes and on the female X chromosome are both up- and down-regulated in the mutant background. The up-regulation of genes may indicate secondary regulatory effects resulting from the loss of MOF. Otherwise, it may indicate that MOF has both activating and repressing effects when promoter-bound.

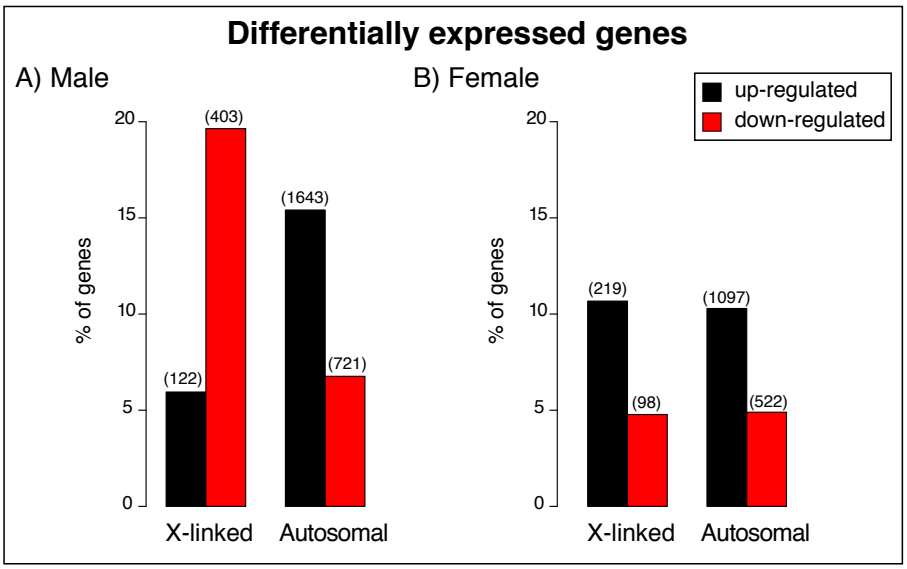


Figure 4.11: Numbers of differentially expressed genes in *mof2* mutants compared wt flies. The barplots show the percentage of X-linked and autosomal genes that are significantly up-regulated (black) or down-regulated (red) in male (A) and female (B) *mof2* mutants. Total numbers of genes are given above each bar.

4.4.2 Down-regulated genes are enriched in MOF-binding and H4K16 acetylation

To distinguish between primary and secondary effects on gene expression, we compared the lists of differentially expressed genes with the pattern of H4K16 acetylation. Since the acetylation mark spreads beyond regions of MOF-binding, we decided to focus our analysis on H4K16ac.

For the X chromosome, we found that 98% of down-regulated genes in the male *mof2* mutant are acetylated in the wt sample; this confirms that H4K16 acetylation along the full length of the gene causes transcriptional activation (p-value=5.71 10^{-5} Fisher test). In females where X chromosomal acetylation is confined to the promoter, only 75% of down-regulated genes in the *mof2* mutant were acetylated. An enrichment test between up- and down-regulated genes showed that the overlap between expression and acetylation was not significant for females. On autosomes, we did not identify any significant trends in acetylation for differentially expressed

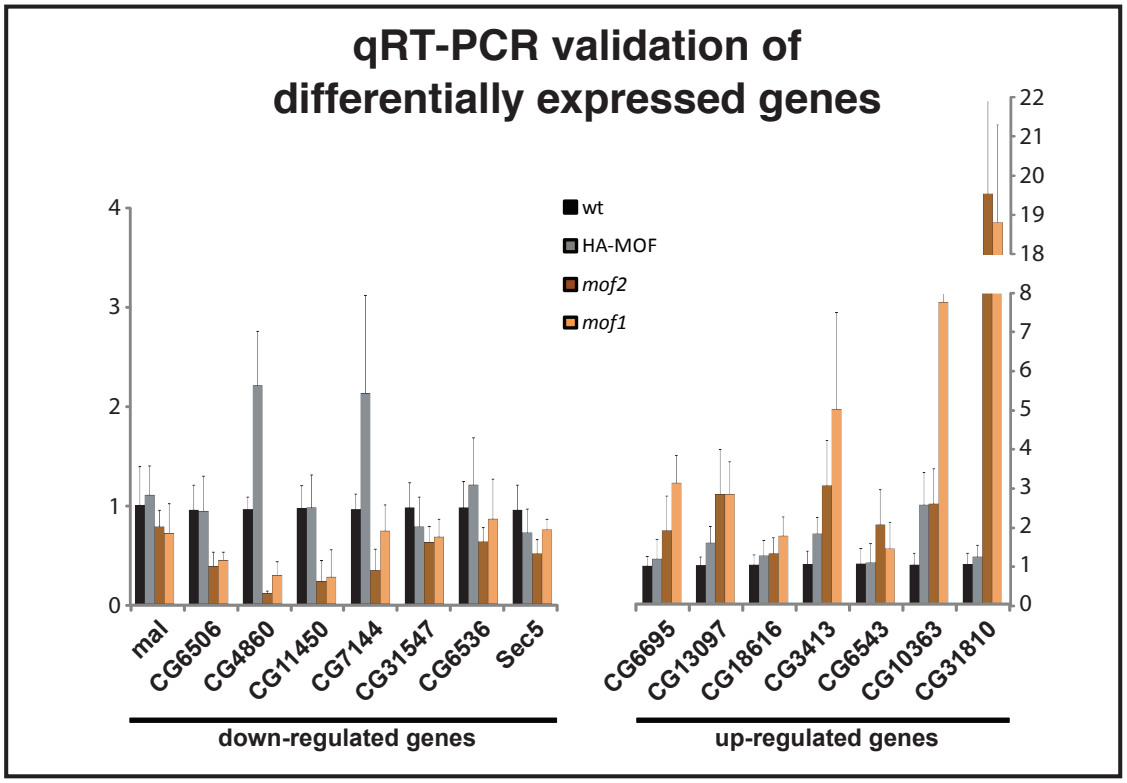


Figure 4.12: qRT-PCR validation of differentially expressed genes. The barplot shows relative expression values of seven up-regulated and eight down-regulated genes that were measured in quadruplicate using qRT-PCR. To avoid biases from global changes in gene expression values, genomic DNA was used as a reference for each gene. Measurements were performed on wt, *mof1* and *mof2* mutant flies as well as a *mof2* mutant complemented with an HA-tagged MOF (indicated in the legend). Gene identifiers are given below. Figure provided by Thomas Conrad.

genes.

However, unexpectedly, we observed a significant proportion of H4K16ac for both up- and down-regulated genes set as compared to non-differentially expressed genes on autosomes and female X chromosome. Frist, we considered the down-regulated gene sets. In agreement with H4K16ac linked to transcription activation, we found a significant proportion of H4K16-acetylated genes among the down-regulated autosomal genes and female X-linked genes in the *mof2* mutants (74% (537), 79% (411) and 77% (75) of down-regulated genes on autosomes in male and female and female X chromosome, respectively; p-value $< 2.2\ 10^{-16}$, p-value $< 2.2\ 10^{-16}$, p-value $= 5\ 10^{-9}$ Fisher test). Second, we considered the up-regulated genes sets. We identified a significant proportion of up-regulated genes in the *mof2* mutants that were

H4K16-acetylated in the wt condition (69% (1140), 78% (855) and 79% (174) of up-regulated genes on autosomes in male and female and female X chromosome, respectively; all p-values $< 2.2\ 10^{-16}$, Fisher test). This last result was surprising, since before most of the up-regulation was attributed to secondary regulatory effects (Kind et al., 2008). This prompted us to evaluate the levels of H4K16ac on *mof2* mutants that will be discussed in section 4.5.

4.5 H4K16 acetylation in the *mof2* mutant

4.5.1 H4K16 acetylation is lost in the gene body and decreases at the promoters in *mof2* mutants

Following the analysis of gene expression changes, we also examined the effect of losing MOF on H4K16ac levels. Since MOF function is closely tied to H4K16ac, we expected to observe a substantial reduction in global acetylation levels in mutant flies. To test this, we generated ChIP-seq data for H4K16ac in male *mof2* mutants, and compared the acetylation pattern with that is observed in wt samples.

Overall, we found pronounced loss of H4K16ac on the X chromosome: only 0.4% of genes (7 out of 1,704 genes) remained fully acetylated in the mutant. (Table 4.2 and Figure 4.13 for gene profile examples).

	X-linked	(%)	Autosomal	(%)	Total	(%)
Total H4K16ac	572	25.66	3,069	26.17	3,641	26.09
- gene body	7	0.31	49	0.42	56	0.4
- promoter	565	25.35	3,020	25.76	3,585	25.69
No H4K16ac	1,657	74.34	8,656	73.83	10,313	73.91
Total	2,229		11,725		13,954	

Table 4.2: Numbers and percentages of H4K16-acetylated X-linked and autosomal genes in the male *mof2* mutant fly.

The loss of acetylation in the gene bodies of X-linked genes contrasted with

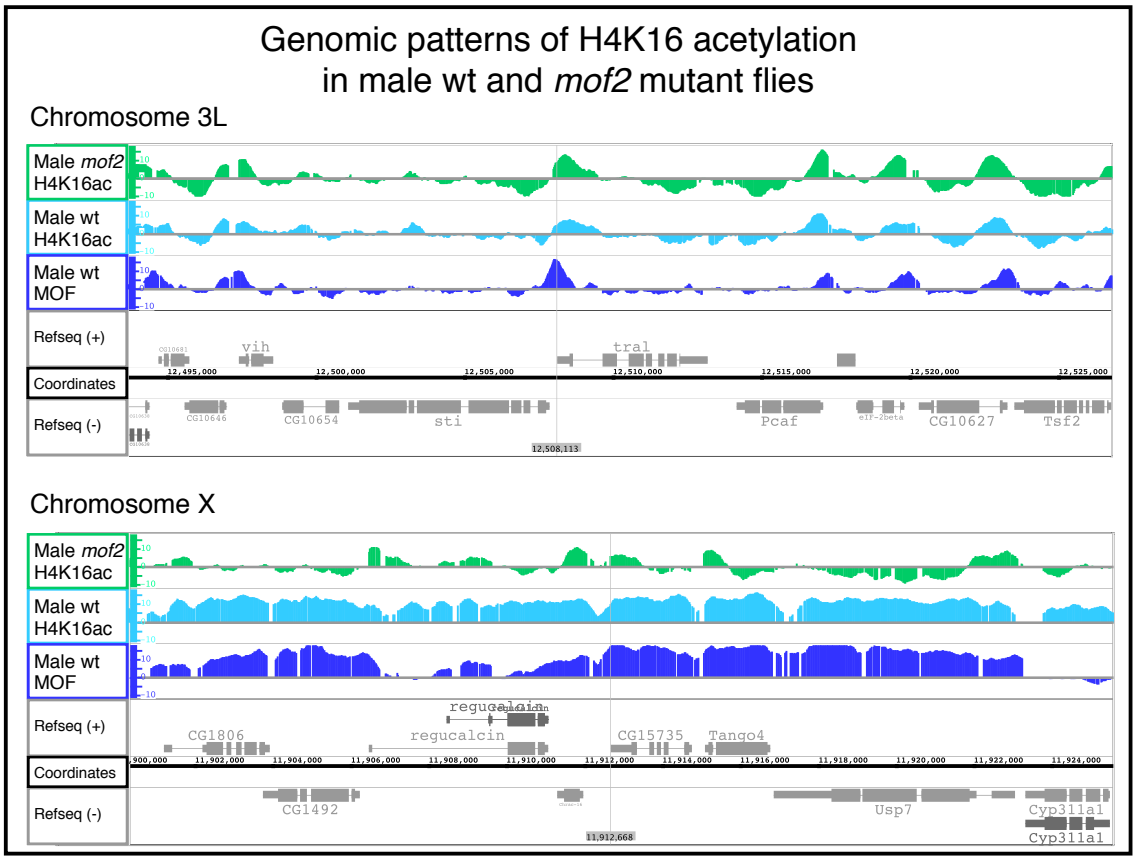


Figure 4.13: Genomic patterns of H4K16 acetylation in male wt and *mof2* mutant flies. Screenshots from the Integrated Genome Browser (Nicol et al., 2009) display the enrichment of H4K16 acetylation in wt males (light blue) compared with *mof2* mutants (green) for a 30kb section of chromosome 3L (top) and a 24kb section of the X chromosome (bottom). We observed a dramatic loss of H4K16ac in the *mof2* mutant. MOF-binding (dark blue) in male wt flies is shown below to illustrate correlation with H4K16ac. Genomic coordinates and the location of annotated Refseq genes on the forward (+) and reverse (-) strands are indicated below.

H4K16ac levels at promoter regions: H4K16ac was still detected for 565 (25%) and 3,020 (25%) of X-linked and autosomal gene promoters, respectively. This indicates a reduction in the amount of acetylation in promoters; however, this result was still very surprising as we expected all H4K16ac marks to be lost in the *mof2* mutant.

We confirmed this observation by performing ChIP followed by qPCR on three X-linked genes and six autosomal genes in the *mof2* mutant (Figure 4.15). As expected acetylation levels were greatly reduced in all gene bodies and promoters tested, but some genes remained acetylated at these reduced levels. Furthermore, polytene staining of *mof2* mutants did not reveal any H4K16ac; however this method

is less sensitive than ChIP (Figure 4.14).

The remaining acetylation could be explained by the presence of another HAT protein that targets H4K16. This could be another enzyme from the MYST protein family that compensates for the loss of MOF-mediated acetylation. It is possible that this yet unidentified enzyme complements MOF function in wt flies, or only becomes active in this way when MOF is removed.

Unfortunately, it was not possible to assess quantitatively the amount by which H4K16ac was reduced at promoters using only ChIP-seq data, since these experiments provide relative measurements of binding or acetylation within a sample. In other words, though a lot of acetylation was lost from male X-linked genes in the *mof2* mutant, we still sequenced this sample to the same depth as the wt. The "extra" reads mapped to regions that were still acetylated in the *mof2* mutant, making it difficult to compare absolute levels of acetylation between samples. Therefore, although the qPCR experiments indicated that H4K16ac was reduced in the mutants, we were not able to demonstrate this with high confidence at a genomic scale.

4.5.2 Up-regulation of autosomal genes

Earlier we observed that acetylated genes were up-regulated in the *mof2* mutant, suggesting a repressive role for MOF on autosomal genes. To examine this further, we compared the expression changes with those found in the *mof1* mutant.

As described earlier, *mof1* contains a MOF mutant with a single point mutation (G691E) in the catalytic domain. The mutation results in a reduced enzymatic activity of 10% of the fully functional protein (Akhtar and Becker, 2001). Preliminary single-gene expression measurements suggested that there is less down-regulation in

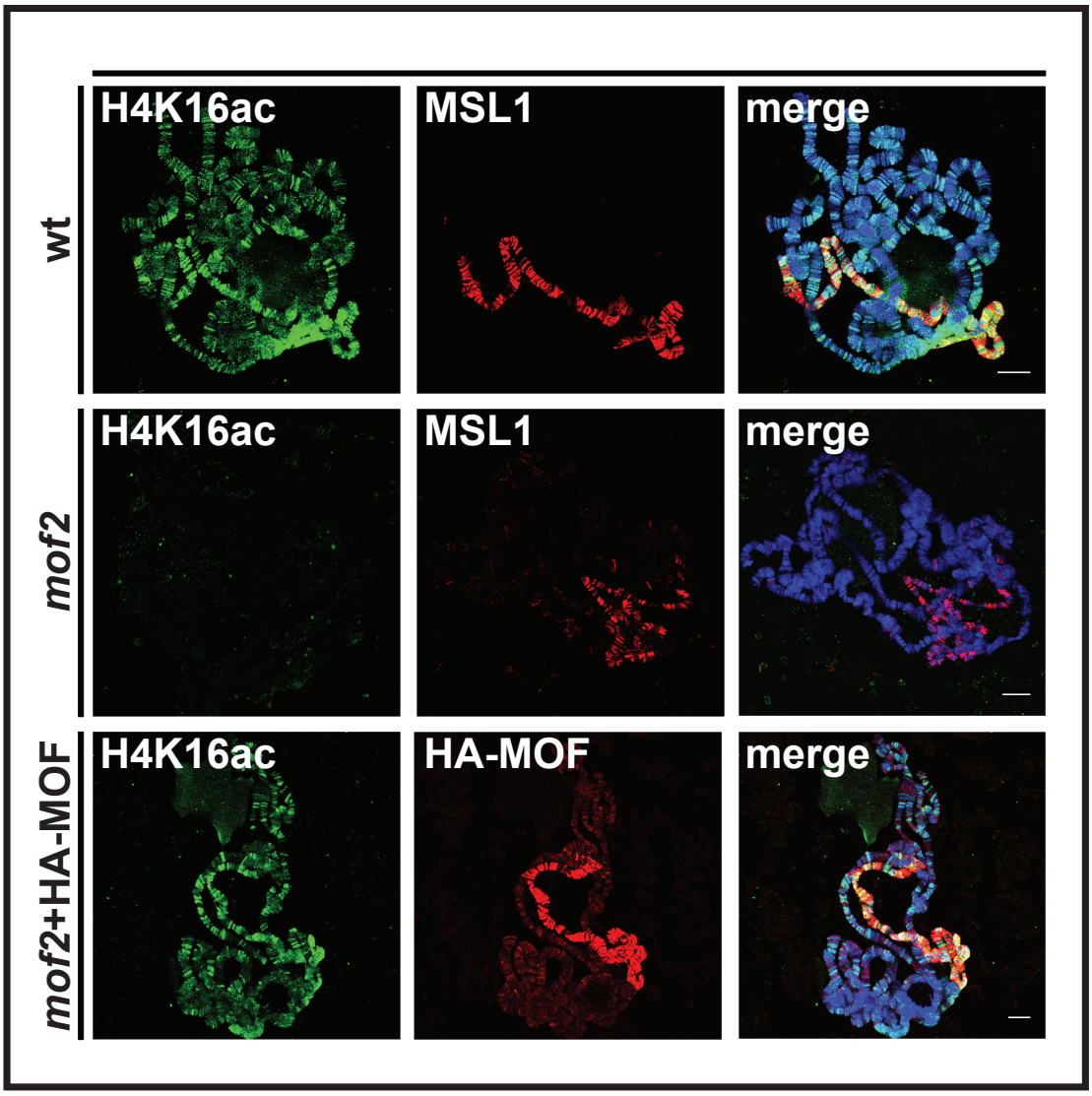


Figure 4.14: Polytene stainings of H4K16ac and MSL1 on chromosomes from wt and *mof2* mutant flies. Chromosomes from salivary gland cells of a wt, *mof2* mutant and *mof2* mutant complemented with a HA-tagged MOF protein (as indicated on the left) were immunostained for H4K16ac, MSL1 and HA-MOF (as indicated inside each image). Figure provided by Thomas Conrad.

the *mof1* mutant compared with *mof2*. However, the amount of up-regulation was similar between the two mutants. This prompted us to examine differences in gene expression changes between the two mutants on a genomic scale.

We observed a large overlap of both up- and down-regulated genes between mutants (Figure 4.16). As with *mof2*, the *mof1* mutant displayed wide-spread down-regulation on the X chromosome, although this effect was moderate in comparison

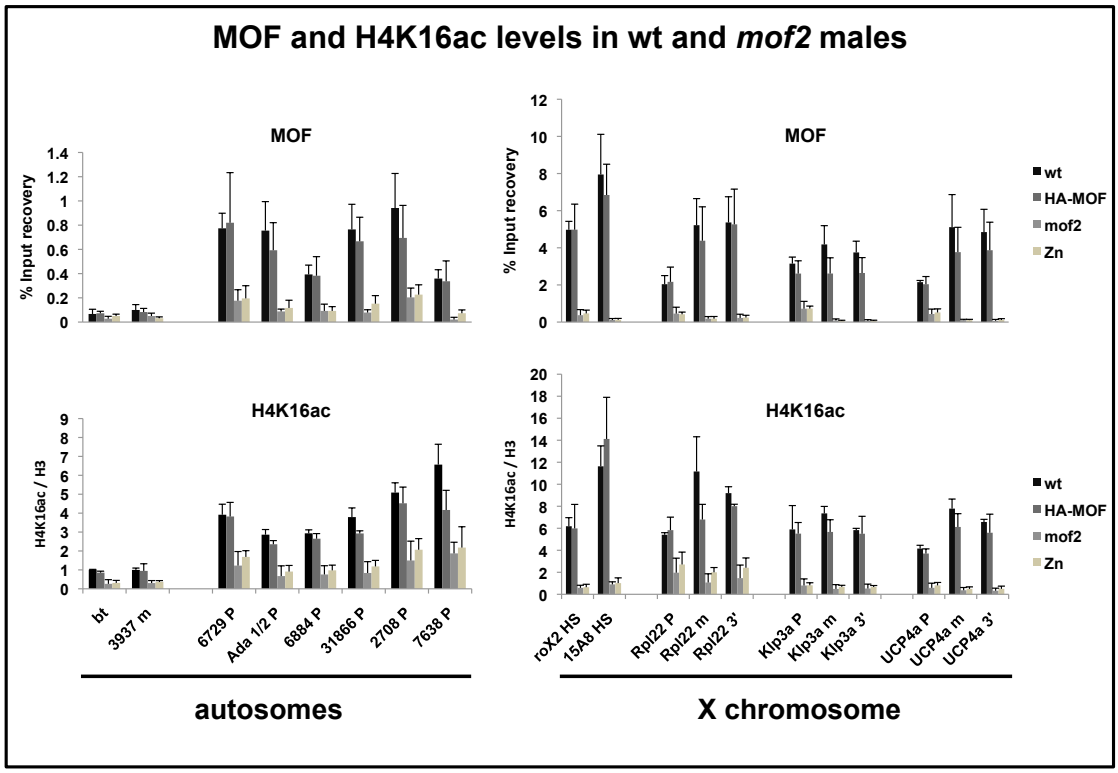


Figure 4.15: qPCR measurements of MOF and H4K16ac levels in male wt and *mof2* mutant flies. The amount of MOF-binding and H4K16ac was measured on six autosomal and three X-linked genes. For the former, only the promoter (P) region was amplified, whereas for the latter the promoter (P), middle (m) and 3'end (3') regions of the genes were amplified. Values for MOF and H4K16ac were normalised against input and H3 IP, respectively. Immunoprecipitations were performed from male wt and *mof2* mutants, and mutants that were transfected with either a HA-tagged MOF protein or a HA-tagged MOF protein lacking the zinc finger domain (ΔZn). All measurements were performed in triplicate. Figure provided by Thomas Conrad.

(10% of X-linked genes down-regulated in *mof1* compared with 20% in *mof2*). In contrast, the amount of differential expression on autosomes is similar for both mutants (13% and 15% of autosomal genes in *mof1* and *mof2*, respectively). Therefore to summarise, similar sets of genes were affected in the two mutants but there were subtle differences: fewer X-linked genes were down-regulated in *mof1* compared with *mof2*, but the effect on autosomal genes was equal.

These observations led us to suggest that MOF might acetylate another target, which then acts as a repressor. This target could be another protein, or a different amino acid residue on the core histone. This other target would be more sensitive to MOF activity than H4K16, since a 90% reduction in HAT activity has the same

effect as a 100% reduction. However, the identity of this target is currently unknown.

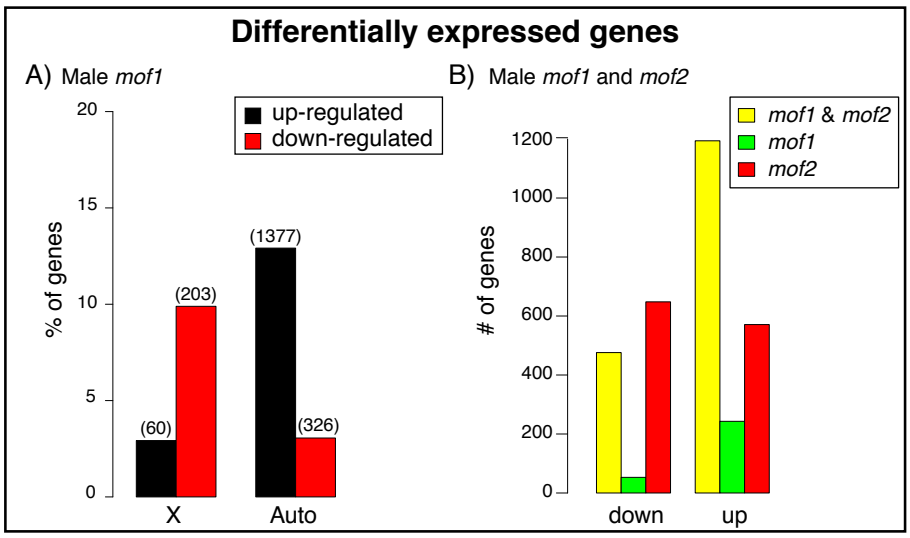


Figure 4.16: Differentially expressed genes in male *mof1* and *mof2* mutants. (A) The barplot shows the percentage of X-linked and autosomal genes (as indicated below) that are significantly up-regulated (black) or down-regulated (red) in male *mof1* mutant flies. Absolute gene numbers are given above each bar. (B) The barplot details the overlap between differentially regulated genes in male *mof1* and *mof2* mutant flies. Genes that were differentially regulated only in *mof1* (green) or *mof2* (red) mutants or that were detected in both (yellow) are indicated.

4.6 Discussion and conclusion

Dosage compensation in *D. melanogaster* is a prime example of transcriptional control at a chromosomal level, as it involves the doubling of expression of the single male X chromosome. Although dosage compensation has been studied genetically and biochemically for over 30 years, advances in genome-scale techniques have led to large advances in our understanding of the system. Recent studies have characterised the full set of proteins involved in dosage compensation (Mendjan et al., 2006), which have been complemented by genome-scale investigations of DNA-binding by these proteins, and the effect they have on gene expression control (Legube et al., 2006; Alekseyenko et al., 2006; Gilfillan et al., 2006; Kind et al., 2008; Gelbart et al., 2009). Despite these advances, there are still many unanswered questions regarding the mechanism of dosage compensation.

In this chapter, we presented the first high-resolution ChIP-seq dataset for the HAT protein MOF, and its corresponding acetylation of histone H4K16. In contrast to previous studies (i) these data were generated from live tissue samples instead of cell lines and (ii) MOF activity was abolished by mutation rather than knocking down by RNAi. We complemented these data with gene expression measurements. The increased resolution of the dataset, in conjunction with the availability of mutants allowed us to examine several aspects of MOF function that were not possible to address before.

In agreement with previous studies from our groups, MOF-binding and H4K16ac follows a bimodal distribution in male X-linked genes – i.e., peaks at the 5' and 3'-ends, with signal throughout the gene body. We observed substantial overlap between MOF-binding and H4K16ac; however acetylation spreads beyond the confines of MOF-binding in the male X chromosome, into neighbouring genes. The observations reinforce the concept that H4K16 acetylation is mediated by transient MOF-binding in the presence of the MSL complex. In contrast to the male X chromosome, MOF-binding and H4K16 acetylation in male autosomes and all female chromosomes were present only at the 5'-end of genes. H4K16 acetylation at promoters were previously associated with transcriptional activity studies (Bell et al., 2007; Schwaiger et al., 2009). However, other studies contradicted these findings, reporting that autosomal promoters were not acetylated (Gelbart et al., 2009). Our results unequivocally confirm that MOF binds at the promoters of autosomal genes in both males and females. Furthermore, these binding events significantly overlap with patterns of H4K16ac.

By assessing H4K16ac in the *mof2* mutants, we confirmed that acetylations are largely lost on the male X chromosome. We also detected reduced levels of H4K16ac in promoter regions, although the amount of reduction was difficult to quantify using ChIP-seq data. These observations were backed by polytene staining and

quantitative ChIP assays for nine genes. Therefore our results establish conclusively, that MOF activity is distributed across the entire genome and is not confined only to the X chromosome.

Discovery of autosomal MOF-binding gave rise to the concept that MOF might play a genome-wide role as a transcriptional regulator beyond just dosage compensation (Kind et al., 2008; Prestel et al., 2010; Hallacli and Akhtar, 2009). A major argument in favour of this view is new evidence showing that MOF also takes part in the Non-Specific Lethal (NSL) complex which preferentially binds to the 5'-end of genes. The complex is present in male and female flies, and its deletion is lethal to both (Mendjan et al., 2006; Raja et al., 2010). A recent study evaluated the association of the NSL complex with chromatin and its transcriptional effect using ChIP-seq and microarray analyses. The results showed that the NSL complex binds uniquely at gene promoters of expressed genes and that there are no biases in terms of X-linked genes (Raja et al., 2010).

Our gene expression analyses presented here provide further support for the view that MOF functions in a global manner. Both *mof1* and *mof2* mutants exhibited large-scale down-regulation of X-linked genes in males. There were also significant numbers of differentially expressed genes on autosomes: but these genes were both up- and down-regulated. The observation—originally made using RNAi-mediated knock-down—was initially explained as resulting from secondary regulatory effects. However our analysis here indicates that both up- and down-regulated genes are likely to be direct MOF targets. This is an unexpected result, since MOF is traditionally considered an activator. A possible explanation is that MOF may target substrates other than H4K16, which then leads to repression of some target genes. This is supported by the fact that the mammalian orthologue of MOF has been shown to acetylate substrates such as histone H4K5 and K8 in addition to K16. MOF is also known to acetylate non-histone targets such as MSL1 and MSL3 within

the MSL complex (Buscaino et al., 2003; Morales et al., 2004). Therefore, loss of acetylation of these other targets might help explain the complex pattern of gene expression changes displayed by the mutants.

Another possible explanation is through the interactions of MOF with the NSL complex. Raja and colleagues recently showed that depletion of NSL subunits by RNAi causes both up- and down-regulation of target genes in males and females. Given that MOF forms part of the NSL complex, it is possible that the transcriptional effects that we observed here arise as a consequence of interactions between MOF and the NSL complex. In future, it will be interesting to examine the overlap between ChIP-seq binding sites for NSL and MOF.

A further benefit of the increased resolution of our analysis is that we were able to determine the preferential binding site for MOF upstream the TSS at around position -100 bp in both male and female cells. It will be interesting to run sequence-motif discovery algorithms with the aim of finding a consensus sequence for MOF-binding. To do so, first we would identify all the MOF-binding peaks, essentially situated at the promoter regions, retrieve the 100 bp sequence regions centered on the peaks, then run the MEME (Bailey and Elkan, 1994) or the NestedMICA (Down and Hubbard, 2005) softwares on these sequences. These programs will return the enriched sequence motifs among the bound regions, which are likely to represent the MOF-binding sequence motif(s). Additionally, it will be of interest to determine whether there are any enrichment for binding by other factors or a combination of histone marks on these regions that would help to recruit the NSL complex.

In summary, our study demonstrated in tissue samples that MOF and H4K16ac are present on both X and autosomal chromosomes of *D. melanogaster*. First, these results reinforce our knowledge that MOF plays a major role as part of the MSL complex in regulating dosage compensation. Second, they support the role of MOF as a transcriptional regulator by binding at the promoter-proximal region of active genes.

Further studies will be necessary to determine the full extent of this regulation. For example, it remains unclear what is the effect of H4K16ac on the chromatin. How does it contribute to the production of double the amount of mRNA from a single template? This and other studies, both at a genome-wide scale and using traditional molecular biology and biochemistry techniques will help us to elucidate the intricate regulation of dosage compensation.

5

Identification of the effect of dosage compensation on Pol II activity

5.1 Introduction

In chapter 4, we investigated the role of the histone acetyltransferease (HAT) MOF as a global transcriptional regulator. We showed that MOF acetylates H4K16 on the male X chromosome directly, and that it also controls H4K16ac levels at the promoters of autosomal genes. Recent studies showed that MOF belongs to two different protein complexes that have distinct binding specificities: the NSL complex directs MOF to promoters, whereas the MSL complex specifies binding towards the 3'-end of genes. Thus the context-dependent activities of MOF are mediated by the membership to these different complexes. As described in earlier chapters, MOF plays a crucial role in dosage compensation as part of the MSL complex; genes on the single-copy X chromosome in males are doubled in expression relative to the diploid autosomes. Although molecular and genomic studies have increased

our knowledge of MSL-complex-binding and function, the mechanism by which it regulates transcription remains unknown. In this chapter, we present a further collaborative study with Dr Asifa Akhtar's laboratory, analysing how Pol II-activity is modified by the presence of MOF-binding and acetylation.

Two competing models have been proposed to explain how dosage compensation might occur: (i) in the "activation model", male X-linked genes are over-expressed owing to binding by the MSL complex, and (ii) in the "sequestration model", autosomal genes are down-regulated as the MSL complex sequesters MOF to the X chromosome. In the first, the MSL complex acts as an activator of MOF, and induces hyper-acetylation of X-linked target genes. In the second, the MSL complex acts as a repressor by sequestering MOF, and its interaction with the X chromosome is not linked to transcriptional activation. Although the two models present quite distinct mechanisms, they both lead to the same outcome in terms of the relative expression levels of X-linked and autosomal genes. This means that it has not been trivial to distinguish between them (reviewed in Straub et al., 2005; Zhang and Oliver, 2007)

The activation model is currently the more generally accepted one, but few studies have provided firm evidence confirming it. The model is largely supported by the observation that the MSL complex interacts extensively with the male X chromosome, and that H4K16 acetylation resulting from the action of MOF is specifically enriched on the chromosome. Kuroda and colleagues presented the first global gene expression analysis after depletion of the MSL complex in male SL2 cell lines. Performing microarray analysis, they showed that the expression of X-linked genes was significantly decreased upon RNAi-mediated knockdown of the MSL2 protein (Hamada et al., 2005). More recently, Kind performed gene-expression analysis after depletion of MOF, MSL1 and MSL3 in SL2 cells (Kind et al., 2008). In each case, there was a large amount of down-regulation among X-linked genes. In another detailed molecular study, expression changes of three MSL-targets were assessed by

RT-PCR; by normalising the RNA signal to genomic DNA, Straub showed that target genes reduced their expression by 60% in MSL-depleted cells (Straub et al., 2005).

In favour of the second model, a study by Birchler showed an up-regulation of autosomal genes upon depletion of the MSL complex, suggesting that MOF is released to activate more genes (Bhadra et al., 1999, 2000, 2005). However, these results must be assessed in context of the results from chapter 4, in which we demonstrated that MOF has a direct repressive effect on some autosomal genes. Following the discovery by Akhtar and colleagues that MOF can form part of two different complexes —the MSL and the NSL complexes (Mendjan et al., 2006; Raja et al., 2010)— Becker and collegues showed that there is competition for MOF between them. Ectopic expression of the MSL2 protein in female Kc cell lines —which allows the formation of the MSL complex— led to decreased interactions between MOF and NSL; this caused greater MOF-association with the X chromosome, and reduced the expression of reporter genes on the autosomes. These results indicated that MOF is expressed at limiting concentrations, and that the MSL and NSL complexes mediate a competition for its activity between the X chromosome and autosomes (Prestel et al., 2010).

As with many scientific debates, true mechanism is likely to be a mix of the above models; however the ongoing discussion highlights that we still do not have a clear picture of how dosage compensation is achieved. In particular we would like to understand how this system achieves a two-fold change in expression.

We are especially interested in understanding the effect that MOF and H4K16 acetylation has on Pol II function. In chapter 4, we showed that in general, autosomal and female X-chromosomal genes are acetylated at the promoter, whereas dosage compensated genes in the male X are acetylated throughout the gene body. Therefore in agreement with previous studies that focused primarily on the X chro-

mosome, these observations have led to suggestions that acetylation in the gene body facilitates transcriptional elongation, rather than enhancing Pol II recruitment to the promoter (Alekseyenko et al., 2006; Gilfillan et al., 2006; Kind et al., 2008).

In this chapter, we combine the data from chapter 4, with additional ChIP-seq measurements of Pol II-binding. In doing so, we aim to assess the effect the MOF and H4K16 on polymerase function.

5.2 Material and methods

5.2.1 Definition of MOF-bound and H4K16-acetylated genes

We used the sets of MOF-bound and H4K16-acetylated genes identified in the previous chapter to compare the effects on Pol II-binding. As described in chapter 4, these genes were subdivided into those with MOF/H4K16ac along the complete gene (full), only in the promoter region (prom) or not at all (not) (Table 4.1 from chapter 4).

To re-iterate the results from chapter 4, we found that in male flies, autosomal genes showed promoter-restricted or no binding, whereas genes on the X chromosome were fully bound. In female flies, the same X-linked genes showed only promoter-restricted binding as did the autosomal genes. Thus, the main difference between males and females is a change of the MOF-binding pattern on X-linked genes.

All the analyses presented in this chapter were performed independently for the MOF-bound genes and the H4K16-acetylated genes. Since MOF-binding and H4K16ac are highly correlated (as demonstrated in chapter 4), the results were very similar for the two gene sets. Here, we present only the results obtained using the set of MOF-bound genes.

5.2.2 Identification of expressed genes and measurement of expression levels

We used the microarray data from chapter 4, section 4.2.2 to assess gene-expression levels in male and female wt samples. Rather than focussing only on differences between wt and mutant flies as in the previous chapter, here we examined the expression levels of all genes in each sample.

To identify expressed genes, we used thresholds defined by the MAS5.0 algorithm implemented in Bioconductor. Briefly, the method compares hybridisation signals from perfectly matching versus mismatching probes on the microarray, and assigns each probeset to one of three categories: (i) "present" in which the matching probes display significantly higher intensity values than mismatching probes; (ii) "absent" in which the signal from matching probes are similar to or lower than those from mismatching probes; and (iii) "marginal" in which the difference between matching and mismatching probes is only marginally significant. We used a very stringent criterion to classify a gene as expressed: all probe sets uniquely mapping to the gene were required to be present in all replicates. For these genes, we then calculated the expression level as the mean log2 expression values from all uniquely mapping probesets in all replicates.

5.2.3 Pol II ChIP-seq data

In addition to the MOF-binding and H4K16ac profiles presented in the previous chapter, we analysed genome-wide binding profiles for Pol II. ChIP-seq experiments were performed using anti-Rpb3 antibodies kindly provided by the Adelman laboratory (Adelman et al., 2005). The antibody recognises the protein irrespective of the phosphorylation state of its carboxy-terminal domain, therefore capturing Pol II during all stages of transcription.

The experiments were performed in duplicates using male and female wt salivary glands from 3^{rd} instar larvae. The sequencing runs benefited from a technological improvement in the Illumina sequencing platform, meaning that more reads were produced per lane compared with the earlier MOF-binding and H4K16ac experiments. The sequencing statistics are displayed in Table B.1.

5.2.4 ChIP-seq data analysis

5.2.4.1 Correlation between replicate samples

To assess the quality of the Pol II ChIP-seq experiments, we checked the correlation between biological replicates. Figure 5.1 shows scatterplots for the raw sequencing data in replicate samples from male and female. There is a high level of correlation ($\rho = 0.74$ for males and $\rho = 0.68$ for females; Spearman correlation), demonstrating the reproducibility of the data.

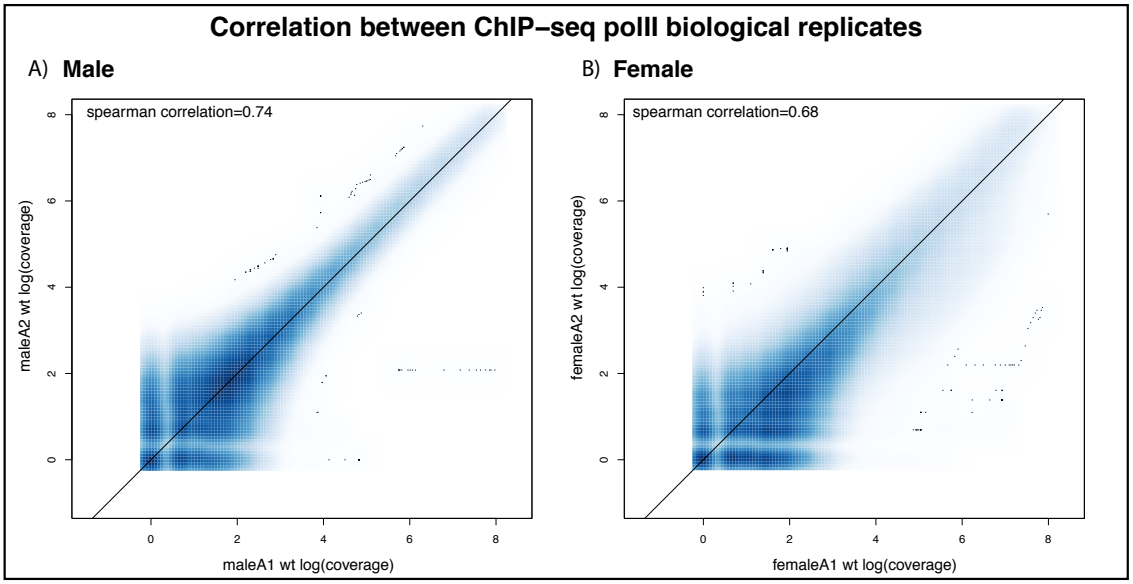


Figure 5.1: Correlation between Pol II ChIP-seq replicates. The smoothed scatterplots show the log (read count) values from the first replicate (x-axis) against log (read count) values from the second replicate (y-axis) for male (A) and female (B) samples. The sequence reads were counted in 25 bp bins along the genome.

5.2.4.2 Data processing to obtain Pol II-binding levels

We processed the Pol II ChIP-seq data as described for the MOF and H4K16ac data in chapter 4, section 4.2.5. Briefly, for each sample we mapped the sequence reads to the *D. melanogaster* genome. We next segmented the genome into 25 bp bins and for each bin we obtained a read count from each experiment. We then used DESeq to normalise the counts to the total number of reads in each sequencing run and to calculate the mean read count between replicates. Finally, we computed the log2 fold change (log2FC) of Pol II IP over input for each bin and smoothed these values using a 400 bp sliding window (Figure 5.3). After determining the null distribution of log2FC values in male and female samples, we calculated the log2FC thresholds to identify genomic regions showing significant Pol II-binding (t = 1.45 and 1.79 for male and female samples, respectively; Figure 5.2).

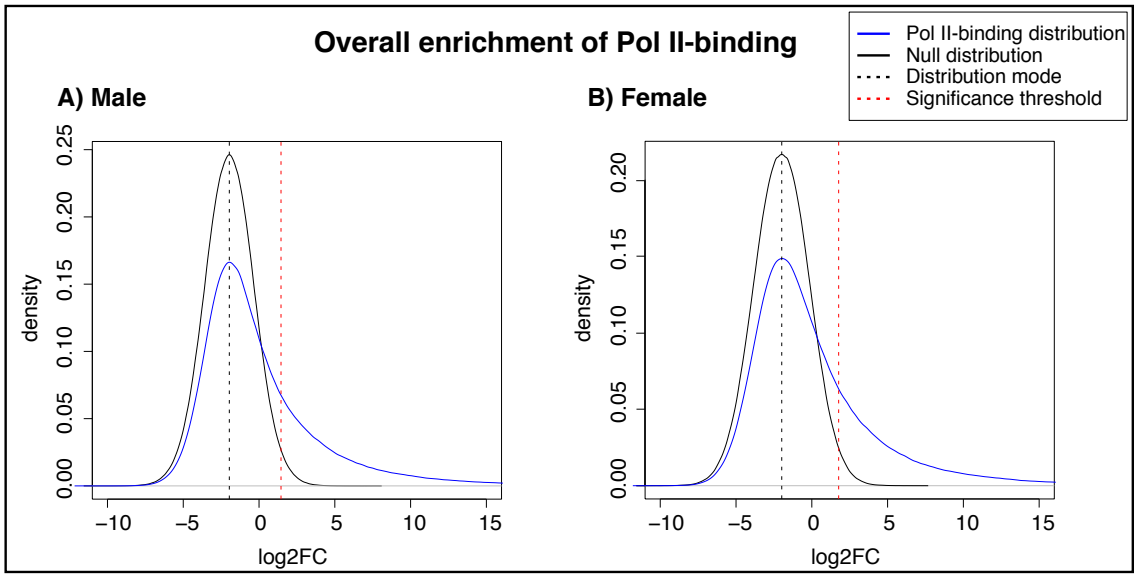


Figure 5.2: Log2FC values for Pol II-binding in males and females. The plots show the smoothed density distributions of log2FC values (Pol II IP over input; blue line) and the corresponding null distributions (black line) and significance thresholds (red dotted line) for Pol II ChIP-seq from male (A) and female (B) samples. The mode of the empirical distribution is indicated by the black dotted line.

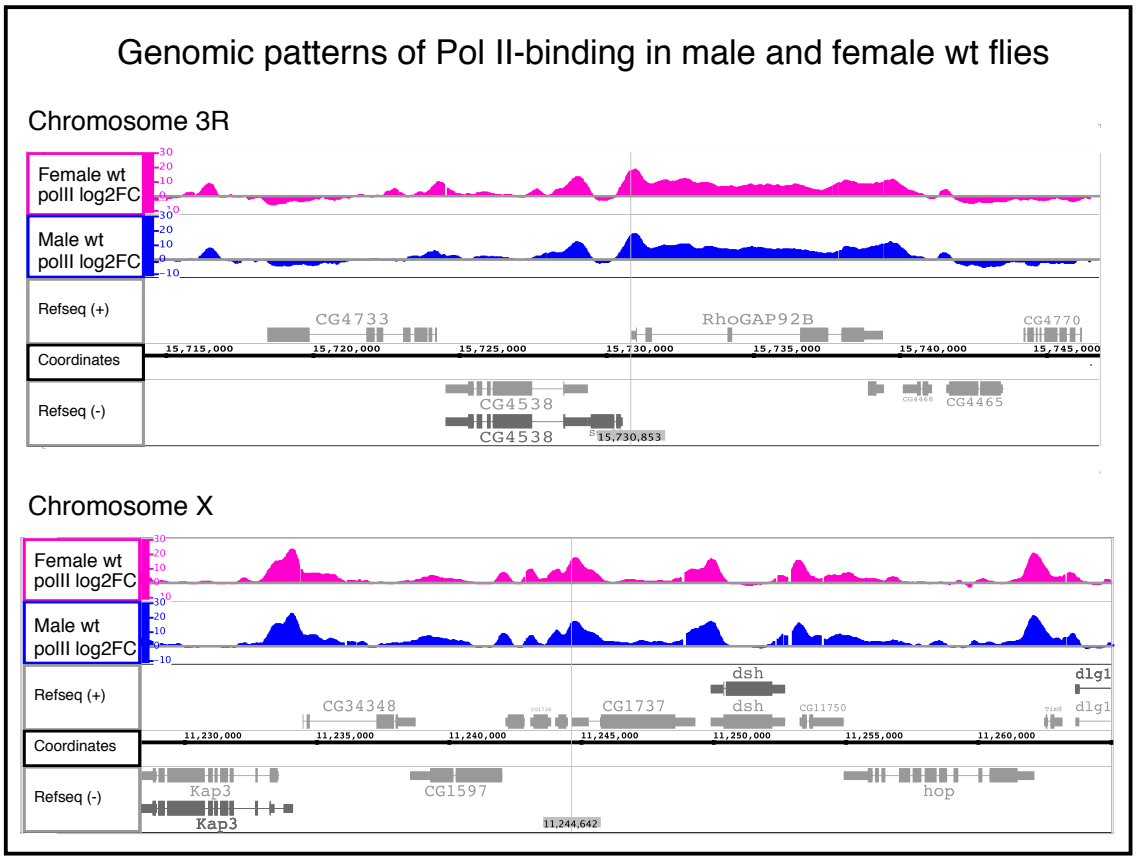


Figure 5.3: Genomic patterns of Pol II-binding. Screenshots of the IGB genome browser (Nicol et al. (2009)) show the enrichment of Pol II-binding in male (blue) and female (pink) salivary glands across a 30 kb sections of chromosomes 3R (top) and X (bottom). Genomic coordinates and the location of annotated Refseq genes on the forward (+) and reverse (-) strand are indicated below.

5.2.4.3 Definition of Pol II-binding levels within genes

To analyse the pattern of Pol II-binding, we computed three measurements for each gene: (i) ProMax, (ii) BodyMedian, and (iii) TailMedian (Figure 5.4). ProMax is the maximum log2FC value in the promoter region which was defined as the 150 bp to either side of the transcription start site (TSS). Thus, ProMax represents the amount of Pol II that is loaded at the TSS of a gene and was used below to evaluate the rate of Pol II release from initiation to elongation. BodyMedian is the median of all log2FC values along the body of a gene. The gene body was defined as the region from 500 bp downstream of the TSS to 500 bp upstream of the 3'-end. This value was only calculated for the 10,484 genes that are longer than 1100 bp (75% of all annotated

Pol II-transcribed loci). BodyMedian represents the amount of Pol II-binding within a gene and thus reflects its level of transcription (see below). TailMedian is the median of all log2FC values in the tail region of a gene which was defined as the region 500 bp upstream of the 3'-end. TailMedian represents the amount of Pol II-binding in the tail of a gene where transcription is terminated. To remove any confounding effects from promoter-binding in neighbouring or overlapping genes on TailMedian values, we removed any genes with a neighbouring gene within 750bp of the termination site (3,772 of all annotated Pol II-transcribed loci; 27%).

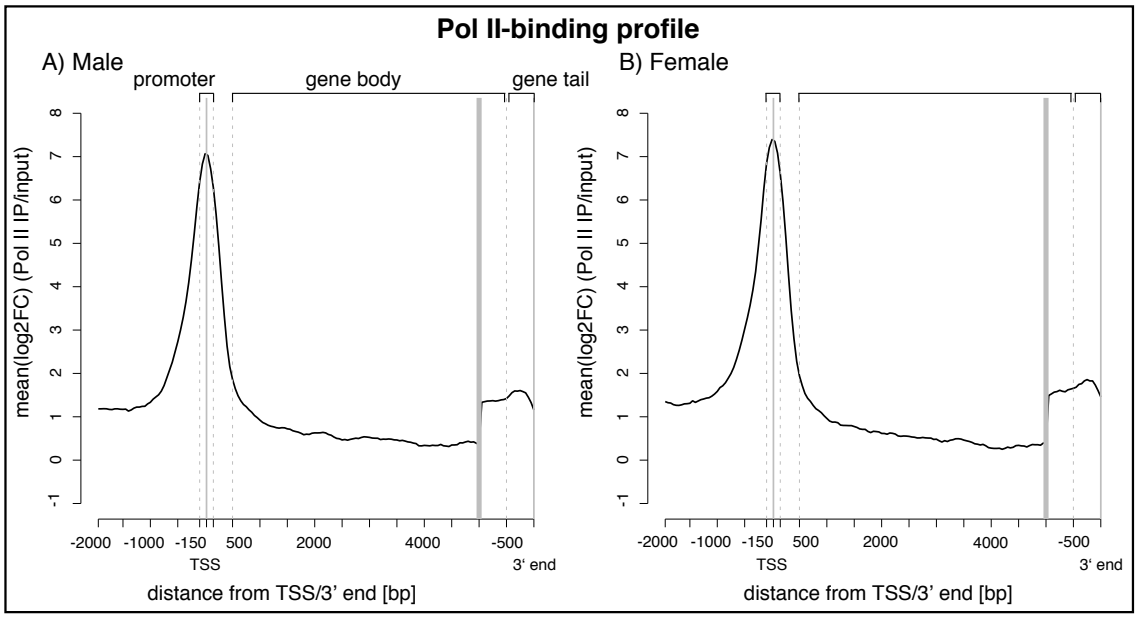


Figure 5.4: Profiles of average Pol II-binding along genes. The plots show the pattern of average Pol II-binding (log2FC) in male (A) and female (B) wt samples. Profiles are shown for the region around the TSS and the gene body (from -2000 bp to 5000 bp), and the 3'-end (from 1000 bp upstream of the transcriptional stop site to the 3'-end). The regions are delimited by grey lines and nucleotide positions are calculated relative to TSS or 3'-end).

5.2.4.4 Calculation of the stalling index

The stalling index is a measure of the relative amount of polymerase that is released from the TSS to proceed into elongation phase. It is calculated as the difference in Pol II-binding between the promoter and the body of a gene. The concept was first introduced by Zeitlinger et al. (2007) in a ChIP-chip study of Pol II-binding in *D. melanogaster*. Here, we defined the stalling index of a gene as the difference between ProMax and BodyMedian.

5.3 Pol II-binding and gene expression

5.3.1 Pol II-binding measurements shows good correlation with microarray expression values

Pol II-binding in gene bodies and expression values obtained from microarray experiments (section 5.2.2) represent two different ways of measuring transcription levels on a genomic scale. On the one hand, ChIP-seq experiments measure actual levels of polymerase-binding during transcription; on the other, microarrays monitor the resulting steady-state levels of mRNA and they are the more generally accepted method for expression level measurements. To assess the agreement between the two data types, we evaluated the correlation between the two data types.

Figure 5.5 shows a scatterplot of BodyMedian and microarray-derived expression values for each gene. There is a high level of correlation between them, confirming that the amount of Pol II-binding in the gene body offers a good estimate of its expression level (ρ=0.61 and ρ=0.63 for males and females respectively, Spearman correlation). We observed the highest correlation for highly expressed genes which benefit from better signal-to-noise ratios in both microarray and ChIP-seq experiments.

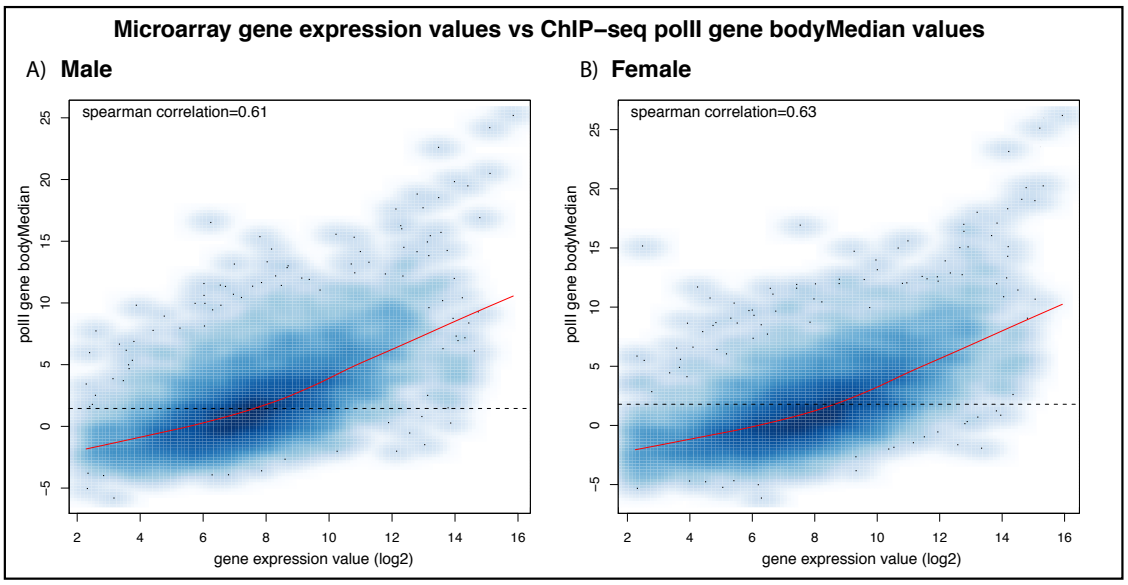


Figure 5.5: Correlation between BodyMedian values and microarray-derived expression values. Smoothed scatterplots show the correlation of microarray-derived expression values (x-axis) with ChIP-seq-derived BodyMedian values (y-axis) for male (A) and female (B) flies. Only genes that were called as present in the microarray analyses are shown.

Gene set	X-linked	Autosomal	Total
Expressed	734	3,733	4,467
Expressed & significant Pol II-binding	254	1,414	1,668

Table 5.1: Number of genes evaluated in this study. Numbers of expressed and Pol II-bound genes on the X chromosome and the autosomes.

5.3.2 Definition of expressed and Pol II-bound genes

To compare the effects of MOF-binding on Pol II activity, we identified the subset of genes that were expressed in both male and female samples. We used stringent criteria based on both microarray and ChIP-seq data to ensure that we analysed only genes for which we detect real Pol II-binding events. Starting with the list of expressed genes as defined by the microarray data, we removed all those that did not have a ProMax value above zero and/or a BodyMedian value above the respective significance threshold (t = 1.45 and 1.79 for male and female datasets, respectively; see section 5.2.4.2 and Figure 5.2). This resulted in a set of 1,668 genes, of which 254 were X-linked and 1,414 were autosomal (Table 5.1). All following analyses were also performed on the full list of microarray-defined expressed genes and gave comparable results.

5.4 Relationship between MOF-binding and Pol II-activity

5.4.1 MOF-bound genes display a higher level of Pol II-binding

Previous studies showed that MOF preferentially binds to expressed genes (Legube et al., 2006; Alekseyenko et al., 2006). Therefore, we compared the BodyMedian values of genes that are bound (both fully and promoter-restricted) or not bound by MOF. The difference in Pol II-binding is significant in both male and female flies (p-values $<2.2\ 10^{-16}$), and the level of Pol II-binding is significantly higher for MOF-bound genes. Since Pol II-binding correlates well with microarray data, our observations agree with the earlier studies. We obtained the same result when the X-linked and autosomal genes were analysed separately (Figure 5.6). We evaluated a possible correlation between the MOF-binding and/or H4K16ac levels and Pol II-binding level. However we did not detect any correlation at either the promoter (X-linked and autosomal genes) or in the gene body regions (X-linked genes). Therefore it appears that it is the presence of MOF and H4K16ac rather than their levels that influences Pol II-binding.

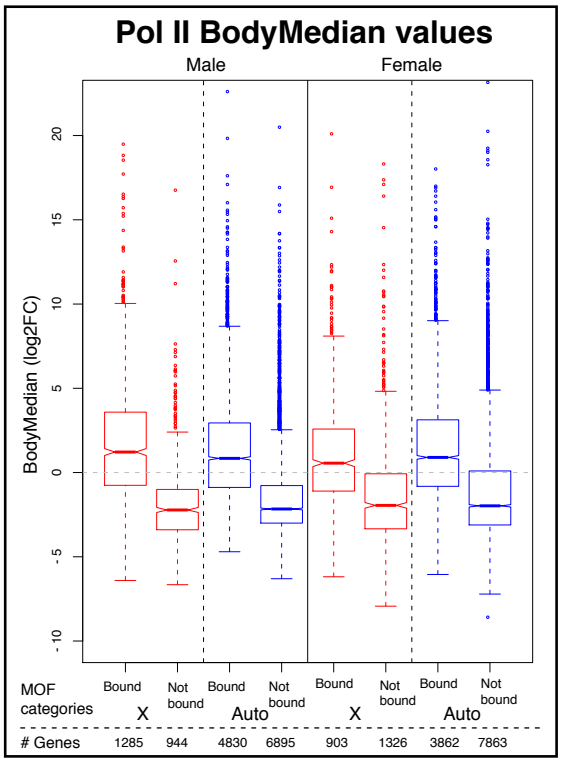


Figure 5.6: Comparison of Pol II-binding in genes that are bound or not bound by MOF. Boxplots display the distribution of BodyMedian values of all MOF-bound genes (full and promoter-restricted binding) and unbound genes (as indicated below) in male (left panel) and female (right panel) samples. Values for X-linked (red) and autosomal (blue) genes are presented separately. The number of genes in each category is given below.

5.4.2 The identification of dosage compensation effects

To examine the effect of dosage compensation on transcriptional activity, we compared the patterns of Pol II-binding between X-linked (i.e., compensated) and autosomal (i.e., non-compensated) genes. Based on the assumption that the majority of genes are equally expressed in male and female wt —as demonstrated by the low number of differentially expressed genes between the two wt samples (see Table B.2, a)—, we should observe equal amounts of Pol II-binding to autosomal genes, but doubled binding to compensated genes on the male X chromosome. Indeed, the input data samples that we use to obtain the log2FC-binding values contain only reads from one X in the male samples whereas they contains reads from two Xs in the female samples. So if the total amount of Pol II-binding on the X-linked genes is

the same in males and females, the log2FC Pol II-binding values will appear doubled in male. To evaluate this hypothesis, for each gene, we subtracted the log2FC Pol II-binding signal in females from that in males. Then, we compared the (male – female) Pol II-binding values between X-linked and autosomal genes.

We classified dosage compensated and uncompensated genes according to the MOF-binding profile in males: compensated genes are fully bound, whereas uncompensated genes are promoter-bound or unbound. Note that there are exceptions to this pattern; there are some promoter and unbound genes on the male X, and fully bound genes on autosomes. However there are too few such examples to be able to draw firm conclusions, and we have excluded them from further analysis.

5.4.3 Pol II displays increased binding in the gene body of dosage-compensated genes

First, we focused on the pattern of Pol II-binding in the gene body since this region should be completely transcribed. For every 25 bp bin in the gene body, we subtracted the Pol II-binding (log2FC values) in females from males. For each gene, we then calculated the median of these differences from all contained bins.

In Figure 5.7, B, we display box-plots of median (male – female) values for dosage compensated (i.e., fully bound by MOF) and uncompensated genes (i.e., promoter- or unbound). There is a significant difference in (male – female) values between them (p-value $<2.2\ 10^{-16}$). Strikingly, this difference is close to 1 (mean = 1.17, confidence interval 95%: [1.05,1.30]), meaning that Pol II-binding is roughly doubled in compensated genes relative to non-compensated ones. It is worth noting that there is no change between promoter-restricted and non-bound genes (p-value = 0.25), confirming that the relative increase of Pol II-binding in the gene body is specific to dosage compensation.

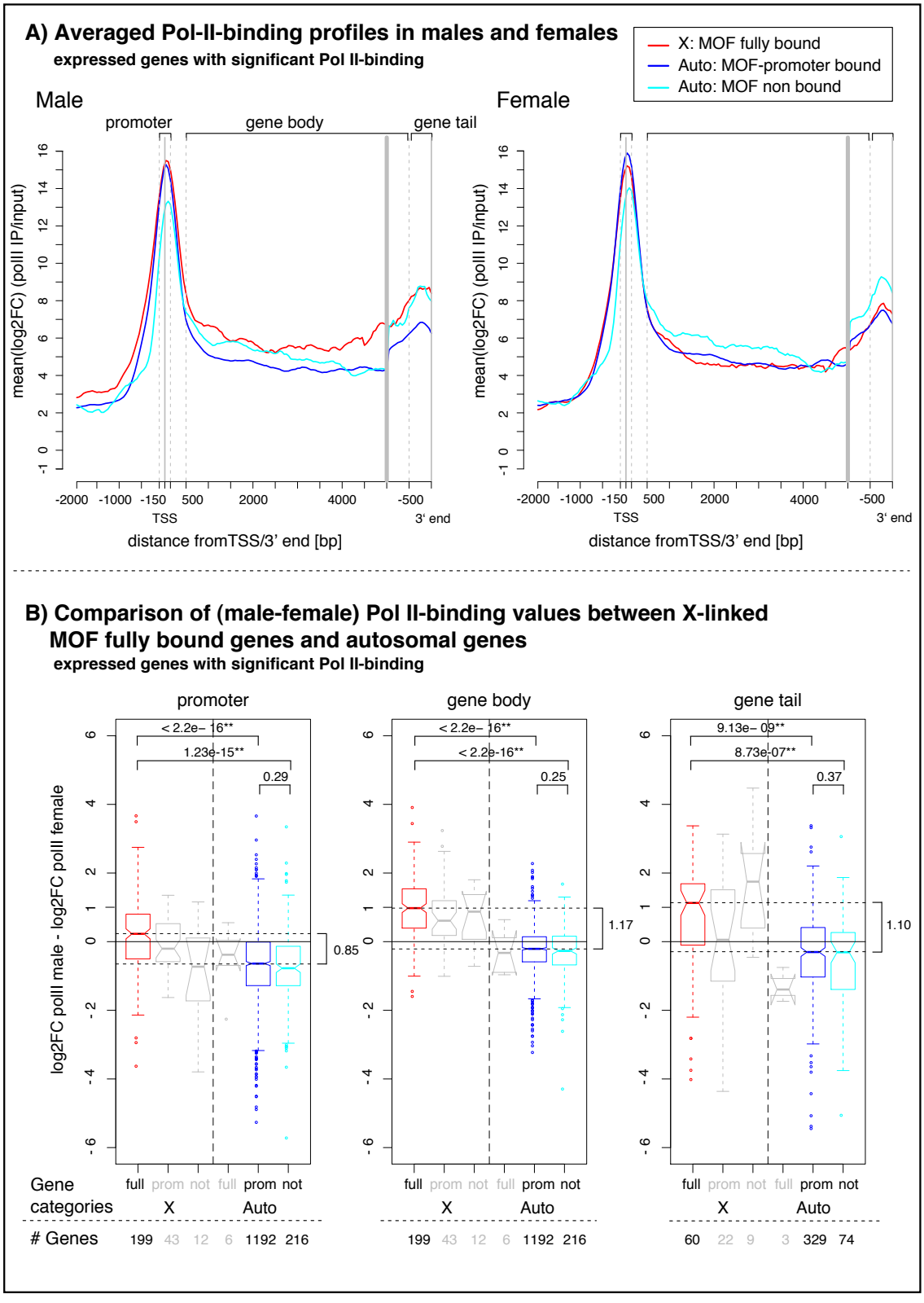


Figure 5.7: Comparison of Pol II-binding in dosage compensated and non-compensated genes. (A) The average Pol II-binding profiles in male (left) and female (right) are shown for X-linked (fully bound by MOF, red) and autosomal genes (promoter-restricted MOF binding, blue; no MOF binding, cyan). Profiles are shown around the TSS and the gene body (from -2000 bp to 5000 bp) as well as for the 3'-end (from 1000 bp upstream of the transcriptional stop site to 3'-end). Regions are delimited by grey lines and positions are calculated relative to TSS or the 3'-end. The three analysed gene regions (promoter, gene body and gene tail) are indicated. (B) The boxplots display the distribution of differences in Pol II-binding between males and femalse in the three gene regions; promoter (left panel), gene body (middle panel) and gene tail (right panel). Compensated and non-compensated genes are colour-coded as in (A). Results for excluded genes (autosomal genes with full binding and X-linked genes with promoter-restricted or no binding) are shown in light grey. P-values using Wilcoxon Rank Sum test are given at the top with brackets indicating the compared lanes. Gene numbers within each category are given below.

5.4.4 Pol II displays increased binding in the 3'-end of dosage-compensated genes

To investigate the influence of MOF on transcriptional activity further, we analysed differences in Pol II-binding at the 3'-end of genes. As described above, we restricted this analysis to genes without neighbouring genes within 750 bp of the 3'-end.

As for the gene bodies, we computed the median (male – female) values in the tail regions of X-linked dosage-compensated and non-compensated genes. We observed a significant difference between these groups of genes (p-value = $9.13\ 10^{-9}$). We did not observe any difference in Pol II-binding between autosomal genes with promoter-restricted or no MOF binding (p-value = 0.37). As for the gene body, these analyses showed that the amount of Pol II-binding at the 3'-end of genes is roughly doubled during dosage compensation (mean difference = 1.10, confidence interval 95%: [0.63,1.57]) (Figure 5.7, B). This suggests that dosage compensation acts equally on transcriptional elongation and termination, since the increase is maintained throughout the gene body and tail.

5.4.5 Pol II displays increased binding in the promoters of dosage-compensated genes

Finally, we evaluated Pol II-binding at the 5'-end of genes. As shown in the average profiles in Figure 5.4, polymerases typically display a peak of binding at promoters, followed by much lower levels of binding in gene bodies. This peak reflects the loading and accumulation of Pol II at promoters before they enter elongation. To determine whether dosage compensation affects Pol II-loading, we compared bind-

ing in the promoters of compensated and non-compensated genes. For this, we calculated (male – female) differences in ProMax values of dosage-compensated and non-compensated genes.

As for gene bodies and tails, we observed significant differences in promoter binding (p-value $< 2.2\ 10^{-16}$, Figure 5.7, B). In contrast to the gene-body analysis however, the average difference was slightly lower than 2-fold (mean difference = 0.85, confidence interval 95%: [0.67,1.02]).

5.4.6 Compensated and uncompensated genes display different levels of Pol II stalling

Given that the observed increase in Pol II-binding was slightly lower at the promoter compared with the gene body, we examined whether dosage compensation influences the release of polymerase from initiation into elongation. For this, we used the stalling index (see 5.2.4.4), which measures the amount of binding in the promoter regions relative to the gene body.

Figure 5.8, B displays the distribution of stalling indices for compensated and non-compensated genes in males and females. There are significant differences in stalling among X-linked and autosomal genes that are promoter bound by MOF (p-value = 0.03 and 0.01, respectively). In both cases, the average stalling index is smaller in males than females, indicating that there is greater Pol II release in males. In contrast, stalling remains the same on genes that are not bound by MOF genes (p-value = 0.52). We also examined whether the difference in stalling is larger among fully bound genes or promoter-restricted genes. As shown in Figure 5.8, C, we confirmed that the effect on stalling is more pronounced in fully bound, dosage-compensated genes.

Since the index is a ratio between Pol II-binding at promoters and gene bodies,

higher values could reflect either increased Pol II loading at the promoter or decreased release of Pol II into the gene body. Unfortunately, because the ChIP-seq data give only relative measures of Pol II-binding within a sample, we could not distinguish between these two scenarios. However, given that dosage-compensation involves MOF binding across the entire gene and we observe a difference in the stalling indexes, we hypothesise that polymerase release might be facilitated by the gene-body acetylation (see Discussion).

Since Zeitlinger et al. (2007) reported in their study that the Pol II stalled genes were enriched for developmental control gene, whereas the ones presenting more uniform Pol II-binding along the genes were enriched for ubiquitously expressed genes, we checked if the dosage compensated or non-compensated set were enriched for any particular function. However a GO enrichment analysis's performed with g:profiler did not allow to detect any particular biological function enriched in these gene sets.

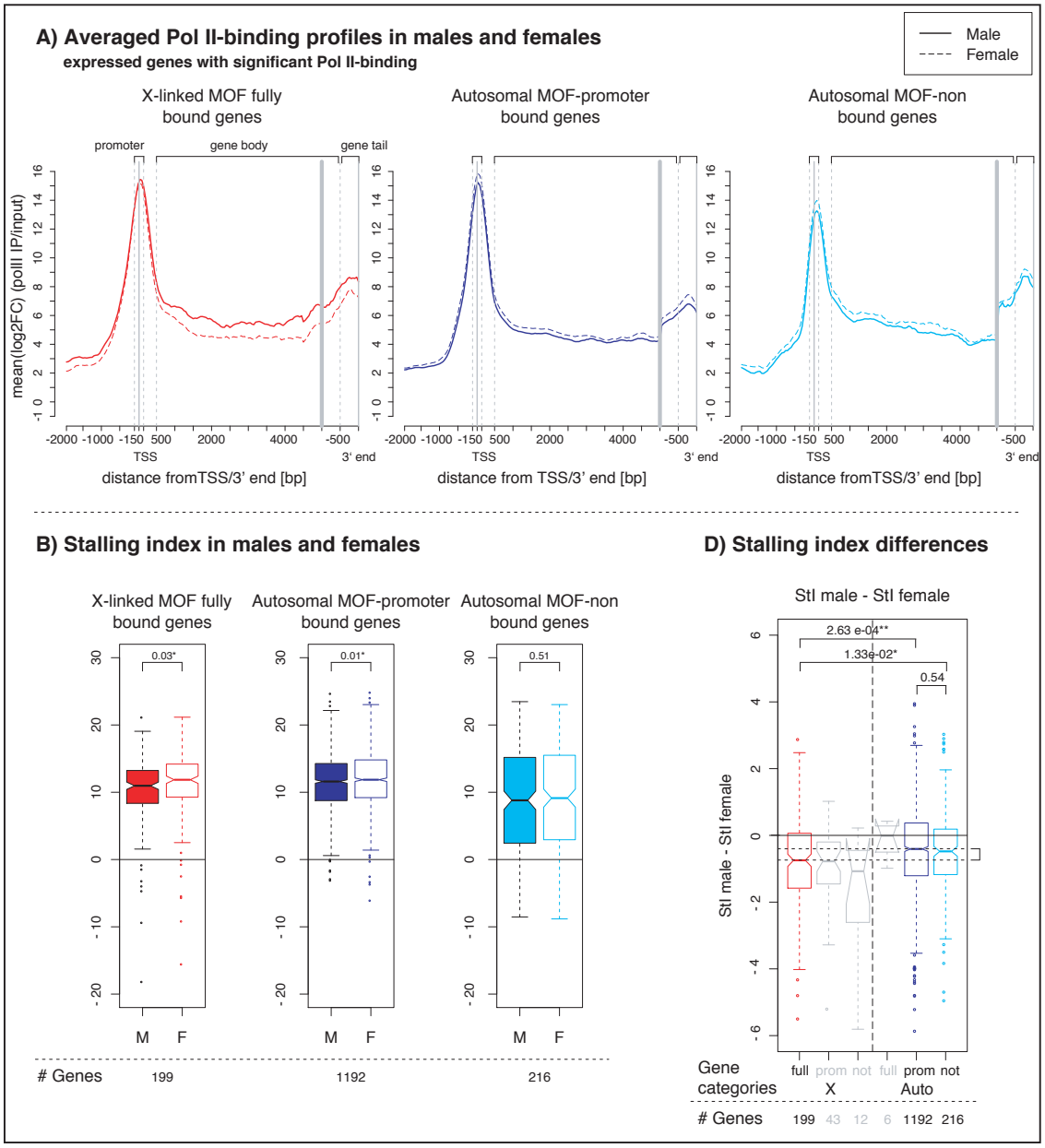


Figure 5.8: Differences in Pol II release in dosage-compensated and non-compensated genes. (A) The average Pol II-binding profiles in dosage-compensated (left panel, X-linked with full MOF binding, red) and non-compensated genes (Autosomal with promoter-restricted [middle panel, blue] or no MOF binding [right panel, cyan]) are shown for male (solid lines) and female (dashed lines). The profiles are shown around the TSS and gene body (from -2000 bp to 5000 bp) as well as the tail region (from 1000 bp upstream of the transcriptional stop site to the 3'-end). Regions are delimited by grey lines and positions are calculated relative to TSS or 3'-end). The three analysed gene regions (promoter, gene body and gene tail) are indicated. (B) The boxplots display the distribution of stalling indices (ProMax- BodyMedian) in males and females for dosage-compensated and non-compensated genes. Categories and colour-coding as in (A). P-values using Wilcoxon Rank Sum test are given. Gene numbers within each category are given below. (C) The boxplots show the differences between male and female in stalling index values for dosage-compensated and non-compensated genes. Categories and colour-coding as in (A). The difference is greater for X-linked genes, indicating that dosage compensation has a significant effect on stalling. P-values using Wilcoxon Rank Sum test are given above, gene numbers in each lane are given below.

5.5 Discussion and conclusion

In this chapter, we examined the effect of MOF-binding and H4K16 acetylation on Pol II function. In particular, we compared the pattern of Pol II-binding among three types of genes found in male and female flies: (i) fully MOF-bound genes, (ii) promoter-bound genes, and (iii) unbound genes.

Since ChIP-seq measurements provide relative levels of binding within a given sample, we calculated all changes in Pol II-binding using the female dataset as an external reference. This was especially convenient for examining Pol II-binding in dosage compensated genes: assuming that genes are similarly expressed in males and females, we would expect to observe equal amounts of Pol II-binding in autosomal genes, and double the amount of binding in compensated genes in the male X chromosome.

In performing these comparisons, we found an almost exact two-fold increase in Pol II-binding in the gene bodies and tails of dosage compensated genes. There was also additional binding in the promoters of these genes, although the increase was less pronounced. This suggested that elongation is indeed elevated when MOF binds the entire gene body; although recruitment of Pol II at promoters is also increased relative to non-compensated genes, this is less pronounced. The stalling index—the ratio between the amount of Pol II-binding at the promoter and in the gene body—confirmed this, showing that stalling is decreased among compensated genes. It would be interesting to have Pol II-binding data of the Ser5P and Ser2P phosphorylation states. In theory, we should have more sensitivity and we would be able to reveal with more confidence the precise differences between the initiation (Ser5P) and elongation (Ser2P) phases.

Of additional interest is the behaviour of Pol II in genes with partial MOF-binding in promoters only. This comparison is less straightforward, since promoter-bound genes in males were also promoter-bound in females: this meant that the nor-

malisation procedure disguised any potential differences between promoter-bound and non-bound genes. Therefore we did not present the results and are currently investigating a better way of assessing this effect.

As discussed earlier, the stalling can be decreased by (i) increasing Pol II-recruitment at the promoter region, or (ii) decrease release of Pol II into the gene body during elongation. Unfortunately because ChIP-seq data provide relative measurements of binding, we were unable to pinpoint the origin of the changes in Pol II activity; PCR experiments measuring absolute binding levels are currently being performed in the Akhtar laboratory, which should provide further insights. Nonetheless, given the above observations, and given the patterns of MOF-binding and acetylation, we can speculate the following: MOF-binding along the entire gene increases both Pol II recruitment at the promoter, and also release into the gene body, resulting in a two-fold increase of expression. It is remarkable that the doubling of Pol II-binding is so precise along the entire length of genes. Further experiments should help reveal how such precise regulation is achieved.

One caveat of this study is that the analysis was restricted to a subset of highly expressed genes (i.e., both expressed by microarray, and displaying Pol II-binding). A major challenge with ChIP-seq experiments for Pol II is that most of the sequence reads map to the highly bound promoter regions, leaving very few reads for the gene bodies. Indeed since Pol II-binding events happen more frequently at the promoter regions than in the gene bodies of transcribed genes, these events are more often caught by the ChIP-seq experiment than Pol II-binding events in the gene bodies. This bias was somewhat ameliorated by sequencing to greater depth; however even with over 30 million reads per ChIP-seq sample, we were unable to detect Pol II-binding in many expressed genes. An alternative approach that we are currently using to overcome this limitation is to perform further ChIP experiments using an antibody that is specific for Pol II in particular phosphorylated states, more precisely

Ser2P to detect polymerase during elongation.

Finally, to differentiate between the "activation" and "sequestration" models that were introduced at the beginning of the chapter, we are currently generating ChIP-seq data for Pol II-binding in a MSL2 deficient background. These experiments will enable us to evaluate changes in Pol II function in males lacking the MSL complex, and allow direct comparisons against the wt male.

In conclusion, the results presented in this chapter have contributed to our understanding of the mechanisms by which MOF regulates transcription. This is the first time that the two-fold increase in Pol II-binding during dosage compensation have been measured. Many questions still remain, and we are currently awaiting additional data that will help us to address them.

6

Discussion and Conclusion

Transcriptional regulation is a complex process involving multiple cis- and trans-factors that act in concert to modulate gene expression. In this thesis, we studied two major contributors to this control: transcription factors (TFs) and histone modifiers.

In the chapters 2 and 3, we explored the role of TFs and their combinatorial usage in determining tissue-specific expression. We analysed the expression of nearly 1,400 TFs across 32 human tissues and identified those that were specifically expressed in a few tissues. The study required the development of a new method, SpeCond, to determine condition-specific expression; the method was implemented as an R package and it is now distributed as part of the Bioconductor software suite. To evaluate the extent of combinatorial regulation among TFs, we integrated protein-protein interaction information with the expression data. This allowed us to identify pairs of interacting TFs that are utilised in a tissue-specific manner. Together, these findings provided insights into how TFs might impose regulatory programmes in different human tissues.

In chapters 4 and 5, we examined the role of a major histone modifier in controlling transcription in flies. The work was performed in collaboration with the Akhtar Group at the Max Planck Institute for Immunobiology and Epigenetics; the experiments were performed by Thomas Conrad of that group, and the computa-

tional analysis was performed by me. We used the dosage compensation system as a model for chromosome-wide transcriptional regulation; in this process, the histone acetyltransferase MOF plays a central role in doubling the expression of genes from the male X chromosome, by targeting histone H4 lysine 16 (H4K16). Using data from high-throughput chromatin-immunoprecipitation experiments (ChIP-seq) and gene expression microarrays, we identified MOF's genomic binding sites in male and female flies. We then assessed the pattern of H4K16 acetylation mediated by MOF's enzymatic activity. Finally, we established that genes that are occupied and acetylated by MOF display a two-fold increase in binding by RNA polymerase II.

This final chapter reviews the main results presented in the thesis, and discusses possible future studies that would expand on these findings.

6.1 Method to identify condition-specific gene expression

The first half of the thesis examined how TFs are utilised in different human tissue types in order to regulate specific gene expression programmes. Determining tissue- or condition-specific expression from microarray data is an important aspect of many genome-scale investigations. However, available statistical techniques did not allow this type of analysis, since most were designed to detect differential expression between a handful of conditions. Therefore, we developed a new method, called SpeCond for Specific Condition, that identifies specifically expressed genes in expression datasets containing measurements across many different conditions.

SpeCond employs a normal mixture model, a well-established statistical approach, to calculate the background distribution of a gene's expression across most conditions. It then scores the gene's specificity in conditions in which it exhibits

unusually high or low expression compared with the background. Since SpeCond does not assume that gene expression values fit a single normal distribution—a common assumption in many types of expression analysis—the method is much more accurate in identifying specifically expressed genes.

We tested SpeCond's performance using a gold standard dataset comprising $\approx 3,500$ specifically and non-specifically expressed genes. In doing so, we showed that SpeCond outperforms other previously published methods. Further, we applied SpeCond to a subset of the SymAtlas microarray dataset to identify specifically expressed genes among 32 normal human tissues. An enrichment analysis of Gene Ontology functions demonstrated that SpeCond helps extract biologically meaningful groups of genes from large microarray datasets. Detailed examination of results revealed that SpeCond also assigns tissues-specificity to previously uncharacterised genes; such information would be useful in investigations of their functions.

As with any method, SpeCond has some limitations. For instance, it requires a minimum of ≈ 10 conditions in the dataset in order to calculate accurate background distributions. In addition, the default parameters set in the distributed software may not be suitable for certain datasets, and users will need to be capable of adjusting these parameters to suit their analyses. Ultimately, the settings will depend on the precise goals of individual studies. However, we have taken care in designing the output so that users can easily understand the method and the effects of different parameter adjustments.

Finally, it would be of great interest to adapt SpeCond to handle RNA-seq data. Although this technique – which measures expression levels using high-throughput sequencing –is currently underused compared with microarrays, it is set to become a major tool for gene expression analysis in the next few years. Analysis methods for identifying differential expression are already available (Anders and Huber, 2010; Robinson et al., 2010), and an approach like SpeCond would be an ideal comple-

ment. However, adapting the method is not trivial; for instance, a statistical model based on normal distributions would be inappropriate. Indeed RNA-seq data provide discrete measurements (rather than continuous ones) following a distribution containing a high number of low values, which are better modeled by a negative binomial distribution (Anders and Huber, 2010).

Nonetheless, we predict that SpeCond will be useful for the foreseeable future. The widespread application of microarrays to biological research over the past decade has generated a vast repository of existing data. Unsurprisingly—given the volume of information—much of these data remain under-analysed and comprehensive meta-studies such as a recent one in human can provide very useful insights (Lukk et al., 2010). SpeCond could be readily applied to these integrated datasets, for example in order for example to identify new gene markers.

6.2 Combinatorial regulation by human TFs

Having developed a way to identify specifically expressed genes, we applied the method to a dataset of TFs in the human genome. We found that different tissues expressed between 3 and 26 TFs specifically; however, since many of them are specifically expressed across several similar tissue types, they are unlikely to provide sufficient specificity on their own.

One way to overcome such limitations is to combine the regulatory activities of multiple TFs; in eukaryotes, this is considered to be an important method to confer a high level of specificity. To explore this, we integrated a dataset of physical protein interactions (PPIs), as a reliable source of evidence for combined TF activity.

Initial examination of the PPI dataset showed that TFs interacted more often with members of the same family. This was not surprising, as several prominent TF families such as the leucine zippers and helix-loop-helix proteins are known to

function as homo- and heterodimers. A possible avenue for further investigation here is the constraint that dimeric interactions have imposed on the evolution of TFs, their binding sites and regulatory influence.

By overlaying the SymAtlas expression dataset on the PPI network, we were able to assess the extent of tissue-specificities of different TF pairs. We identified a large number of TF pairs that are specifically expressed. About 40% of these pairs involved at least one TF that is itself specific; however of particular interest were pairs in which each TF was non-specific, but were co-expressed in only a few tissues. Given these observations, we tested the ability to predict the overall expression pattern of genes, given the presence or absence of individual TFs or pairs of TFs. Our linear modelling approach showed that TF pairs provided greater predictive power than individual TFs, but the combined use of both types of information gave the best predictions. Finally, we examined the use of specific TF pairs in different tissues; we highlighted common sets of TF pairs in related tissues, as well as those that were entirely unique.

It is important to note that a recent study published by Ravasi and colleagues presented the first large-scale analysis of TFs combinations in human and mouse Ravasi et al. (2010). The work presented in this thesis was started before this publication; though this was an unfortunate coincidence, the fact that many observations are shared between the two studies highlights the robustness of the results.

There are several interesting follow-up analyses that could be done. We focused here on pairwise TF interactions in order to simplify the analysis. Crystal structures like that of the enhanceosome depicts the combinatorial regulation by four different TFs; it may be useful to incorporate information about higher order TF combinations into future analyses. On a related note, TFs are part of a much wider network of PPIs including co-factors and histone modifiers that directly impact on polymerase function. Clearly, identifying the relevant protein interactions and interpreting their

regulatory influence in mammalian systems is an enormous task; however, the best approach for such a project is by examining particular systems (Amit et al., 2009).

6.3 Revealing the effects of the MOF histone acetyltransferase

The second half of the thesis studied the effects of histone modifications on transcriptional regulation in *D. melanogaster*, using dosage compensation as a model. Dosage compensation in flies is achieved through the regulatory activity of the Dosage Compensation Complex; an important component of the complex is the acetyltransferase protein MOF, which targets lysine K16 on histone H4.

Early studies using polytene staining revealed a specific binding of the DCC to the male X chromosome (Kelley et al., 1999) that was accompanied by large amounts of H4K16 acetylation (Bone et al., 1994). This pattern of binding and acetylation was later confirmed at a higher resolution by ChIP-chip studies (Alekseyenko et al., 2006; Gilfillan et al., 2006; Kind et al., 2008; Gelbart et al., 2009). Though dosage compensation has been studied at a molecular level for $\approx$ 30 years, the advent of genome-scale experimental techniques has accelerated the rate of new discoveries in this system. A particularly surprising finding by Kind *et al.*, was the observation that MOF functioned differently on the male X chromosome, where it is involved in dosage compensation, and on autosomes where it behaves like a global transcriptional regulator. Subsequent studies showed that this context-dependent behaviour is mediated by two different protein complexes MSL and NSL which compete for MOF-binding.

To understand the functional implications of this dual regulatory behaviour, we collaborated with the Akhtar Group in Freiburg to analyse data from a compre-

hensive set of ChIP-seq and microarray experiments describing H4K16 acetylation, MOF- and Pol II-binding. We used these data to answer these questions: (i) what is the relationship between MOF-binding and H4K16 acetylation; (ii) what is the effect of MOF-binding and acetylation on gene expression; and (iii) what is the effect of MOF-binding and acetylation on Pol II activity.

6.3.1 Computational challenges in analysing ChIP-seq data

Much of the work presented here relied on the analysis of ChIP-seq data. Since the technology is still relatively new, there are still no established standards for processing these data. One particular challenge was the identification of binding sites. Most current ChIP-seq analysis methods were designed for peak calling (such as that observed in TF-binding), rather than for detecting extended chromatin marks or binding domains (reviewed in Pepke et al., 2009). Therefore we had to develop new computational protocols to handle these types of ChIP-seq data.

The analysis pipeline we developed is based on the DESeq algorithm, which was originally designed for measuring differential gene expression in RNA-seq data (Anders and Huber, 2010). We customised the method by applying it to 25 bp bins rather than entire genes in order to calculate signal from the ChIP sample relative to the control (log2FC values). We also applied a method for calculating a threshold of statistically significant binding originally developed for ChIP-chip data (Schwartz et al., 2006) using the continuous log2FC values computed earlier as input.

A natural extension of the analysis pipeline would be to allow direct comparisons between different types of biological conditions. Currently, we can only calculate log2 fold change values relative to other bins within a given condition. This meant that we had to calculate different significance thresholds for each sample type (e.g., Pol II-binding in males and females), and we had to use rather convoluted approaches to compare binding patterns between samples (e.g., subtracting the Pol II-binding

signal in females from that in males). An improved normalisation based on the distributions of log2FC values would enable us to define a single significance threshold across all samples, and would also allow direct comparison of binding levels.

6.3.2 Analysis of MOF-binding and H4K16 acetylation

The ChIP-seq data generated by our collaborators are the first high-resolution binding profiles for the MOF histone acetyltransferase in *D. melanogaster*. These experiments provided several important advantages over earlier studies, leading to new insights that were previously inaccessible. These include: (i) the use of next-generation sequencing technologies, instead of microarrays, which allowed us to examine the system at an unprecedented level of resolution; (ii) the use of living tissues, rather than immortalised cell line, which allowed us to test the system in vivo; and (iii) the use of MOF mutants, instead of RNAi-mediated knock-down, which allowed us to evaluate the effects following total loss of MOF function.

As shown previously, we observed the dual regulatory behaviour of MOF in male and female samples. We identified $\approx 98\%$ of the binding sites previously reported by Kind et al. (2008), and indeed, we were able to detect several thousand new binding sites.

By comparing the patterns of MOF-binding and H4K16ac, we demonstrated that MOF functions as a global histone modifier, acting on both the X chromosome and autosomes. We showed that binding and acetylation are extremely correlated in autosomes, but that acetylation extends beyond regions of MOF-binding. The effect of removing MOF was dramatic, leading to almost complete loss of H4K16ac on the male X chromosome, and a substantial reduction of the mark on autosomes.

These observations settle a minor controversy surrounding the role of MOF as a global regulator. Our results are consistent with the original findings by Kind *et al.*, who originally reported the agreement in MOF-binding and H4K16ac on autosomes

(but did not measure acetylation levels upon MOF-depletion). Our results contradict the claim by Gelbart et al. (2009), who reported that they could not detect any reduction in H4K16 levels upon MOF knockdown and therefore argued that MOF could not be a global regulator. Interestingly, both our ChIP-seq and qPCR data show that acetylation is not completely lost on autosomes, suggesting that there is residual enzymatic activity from an as yet uncharacterised acetyltransferase protein.

6.3.3 Impact of MOF-binding on gene expression

To assess the impact of MOF-mediated acetylation on gene expression, we integrated microarray data from wt and mutant flies. As expected, we detected a large amount of down-regulation among X-linked genes in the mutant, which most likely resulted from the loss of the dosage compensation mechanism. However, we also observed substantial down- as well as up-regulation among autosomal genes; this was unexpected, as it seemed to contradict the near-universal view that histone acetylation leads to transcriptional activation.

Further analysis showed that most of the up-regulated genes were bound and acetylated by MOF in wild type, ruling out the possibility that these changes were due to secondary effects. Our current working hypothesis is that MOF also acetylates another substrate, which then leads to repression of some target genes. MOF is known to target non-histone proteins such as MSL1 and MSL3 within the dosage compensation complex (Buscaino et al., 2003; Morales et al., 2004), and the mammalian orthologue of MOF acetylates histone H4K5 and K8 in addition to K16.

In order to establish conclusively, MOF as a global transcriptional regulator, and to explain its behaviour, it will be important to identify the full complement of enzymatic targets. It will also be crucial to understand how MOF's function is modulated by its different interaction partners.

6.4 Impact of MOF-binding on Pol II-activity

Finally by integrating ChIP-seq data for Pol II-binding in wt and mutant flies, we analysed MOF's effect on the transcriptional process itself. For this, we compared patterns of Pol II-binding in three types of genes: (i) fully bound by MOF; (ii) promoter-bound by MOF and (iii) unbound.

It was important to use Pol II-binding data from female samples as an external reference; since the analysis was based on the assumption that if genes are similarly expressed in males and females, we should find double the amount of Pol II-binding in compensated genes compared with uncompensated ones. This strategy allowed us to identify a clear two-fold increase in Pol II-binding in the gene bodies and tails of compensated genes. We also observed increased binding in the promoter; however this increase was less than two-fold. In order to understand which stages of transcription are affected by MOF, we calculated a stalling index that measures the ratio of Pol II-binding between the promoter and gene body. Compensated genes showed decreased stalling, indicating that Pol II is released from initiation into elongation at a faster rate than in uncompensated genes.

Taken together, these results allowed us to propose the following model: MOF-binding along the entire gene increases both Pol II recruitment and release into the gene body, resulting in a two-fold increase of actively transcribing Pol II. We speculate that H4K16 acetylation contributes to this mechanism; however at this stage it is not clear whether this is a direct result of a change in chromatin structure, or the mark helps recruit additional transcription activators.

There are plans to perform several further analyses to strengthen and improve these results. First, we are currently performing additional ChIP experiments for Pol II in the mof2 mutant. These data will allow us to establish firmly the causal effect of MOF on Pol II-binding. In addition since the current Pol II-binding data do not allow us to detect binding in many lowly expressed genes, we are repeating

these experiments using an antibody for the Ser2P phosphorylation mark on the CTD tail.

Several questions remain unanswered. As ChIP-seq data measure only relative levels of binding within the sample, it is not possible to determine whether the differences in Pol II-binding originate from enrichments in one or the other sex. To clarify this, we will perform small-scale ChIP experiments using genomic DNA as the reference; this will provide information about absolute Pol II-binding levels. This will allow us to determine unequivocally the origin of the fold-change observed between compensated and uncompensated genes. Taken together, these new data should confirm the effect of MOF-binding and histone acetylation on Pol-II activity.

6.5 Concluding remarks

In conclusion, this thesis presented a computational study of transcriptional regulation on a genomic scale. We analysed regulation at two different levels: (i) at the level of TFs; and (ii) at the level of histone modifications. In the first, we identified groups of human TFs that drive tissue-specific gene expression patterns, and demonstrated the importance of combinatorial regulation. In the second, we assessed the influence of the histone acetyltransferase MOF on the global expression programme of flies, and measured directly the effect it has on Pol II function. Together, these analyses contribute to a better understanding of how transcription is regulated in a complex organism, and reveal new findings that provide the basis for future investigations.

Bibliography

Adams, M. D. (2000). The Genome Sequence of Drosophila melanogaster. *Science*, 287(5461):2185–2195.

Adelman, K., Marr, M. T., Werner, J., Saunders, A., Ni, Z., Andrulis, E. D., and Lis, J. T. (2005). Efficient release from promoter-proximal stall sites requires transcript cleavage factor TFIIS. *Molecular cell*, 17(1):103–12.

Akhtar, A. and Becker, P. B. (2000). Activation of transcription through histone H4 acetylation by MOF, an acetyltransferase essential for dosage compensation in Drosophila. *Molecular cell*, 5(2):367–75.

Akhtar, A. and Becker, P. B. (2001). The histone H4 acetyltransferase MOF uses a C2HC zinc finger for substrate recognition. *EMBO reports*, 2(2):113–8.

Alekseyenko, A. A., Larschan, E., Lai, W. R., Park, P. J., and Kuroda, M. I. (2006). High-resolution ChIP – chip analysis reveals that the Drosophila MSL complex selectively identifies active genes on the male X chromosome. *Genes & Development*, pages 848–857.

Alekseyenko, A. A., Peng, S., Larschan, E., Gorchakov, A. A., Lee, O.-K., Kharchenko, P., McGrath, S. D., Wang, C. I., Mardis, E. R., Park, P. J., and Kuroda, M. I. (2008). A sequence motif within chromatin entry sites directs MSL establishment on the Drosophila X chromosome. *Cell*, 134(4):599–609.

Allis, D., Jenuwein, T., Reinberg, D., and Caparros, M.-L., editors (2007). *Epigenetics*. Cold Spring Harbor Laboratory Press.

Amit, I., Garber, M., Chevrier, N., Leite, A. P., Eisenhaure, T., Guttman, M., Grenier, J. K., Li, W., Zuk, O., Schubert, L. A., Birditt, B., Shay, T., Goren, A., Zhang, X., Deering, R., Mcdonald, R. C., Cabili, M., Bernstein, B. E., Rinn, J. L., Meissner, A., Root, D. E., Hacohen, N., and Regev, A. (2009). Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. *Science*, 326(5950):257–263.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106.

Bailey, T. L. and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, AAAI Press, Menlo Park, California*, pages 28–36.

Bannister, A. J. and Kouzarides, T. (2005). Reversing histone methylation. *Nature*, 436(7054):1103–6.

Bannister, A. J., Zegerman, P., Partridge, J. F., Miska, E. A., Thomas, J. O., Allshire, R. C., and Kouzarides, T. (2001). Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature*, 410(6824):120–4.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–37.

Bell, O., Conrad, T., Kind, J., Wirbelauer, C., Akhtar, A., and Schübeler, D. (2008). Transcription-coupled methylation of histone H3 at lysine 36 regulates

dosage compensation by enhancing recruitment of the MSL complex in Drosophila melanogaster. *Molecular and cellular biology*, 28(10):3401–9.

Bell, O., Wirbelauer, C., Hild, M., Scharf, A., Schwaiger, M., MacAlpine, D., Zilbermann, F., van Leeuwen, F., Bell, S., Imhof, A., Garza, D., Peters, A., and D., S. (2007). Localized H3K36 methylation states define histone H4K16 acetylation during transcriptional elongation in Drosophila. *EMBO J.*, 24(26):4974–84.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29:1165–1188.

Bernstein, B. E., Mikkelsen, T. S., Xie, X., Kamal, M., Huebert, D. J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., Jaenisch, R., Wagschal, A., Feil, R., Schreiber, S. L., and Lander, E. S. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, 125(2):315–26.

Bertone, P., Stolc, V., Royce, T. E., Rozowsky, J. S., Urban, A. E., Zhu, X., Rinn, J. L., Tongprasit, W., Samanta, M., Weissman, S., Gerstein, M., and Snyder, M. (2004). Global identification of human transcribed sequences with genome tiling arrays. *Science*, 306(5705):2242–6.

Bhadra, M. P., Bhadra, U., Kundu, J., and Birchler, J. A. (2005). Gene expression analysis of the function of the male-specific lethal complex in Drosophila. *Genetics*, 169(4):2061–74.

Bhadra, U., Pal-Bhadra, M., and Birchler, J. A. (1999). Role of the male specific lethal (msl) genes in modifying the effects of sex chromosomal dosage in Drosophila. *Genetics*, 152(1):249–68.

Bhadra, U., Pal-Bhadra, M., and Birchler, J. A. (2000). Histone acetylation and gene expression analysis of sex lethal mutants in Drosophila. *Genetics*, 155(2):753–63.

Blalock, E. M., Geddes, J. W., Chen, K. C., Porter, N. M., Markesbery, W. R., and Landfield, P. W. (2004). Incipient alzheimer ' s disease : Microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *PNAS*, 101:2174–2178.

Boehm, A. K., Saunders, A., Werner, J., and Lis, J. T. (2003). Transcription Factor and Polymerase Recruitment , Modification , and Movement on dhsp70 In Vivo in the Minutes following Heat Shock. *Society*, 23(21):7628–7637.

Boggaram, V. (2009). Thyroid transcription factor-1 (TTF-1/Nkx2.1/TITF1) gene regulation in the lung. *Clinical science*, 116(1):27–35.

Bone, J. R., Lavender, J., Richman, R., Palmer, M. J., Turner, B. M., and Kuroda, M. I. (1994). Acetylated histone H4 on the male X chromosome is associated with dosage compensation in Drosophila. *Genes & development*, 8(1):96–104.

Bonn, S. and Furlong, E. (2008). cis-Regulatory networks during development: a view of Drosophila. *Curr Opin Genet Dev*, 18(6):513–20.

Bottomley, M. J. (2004). Structures of protein domains that create or recognize histone modifications. *EMBO reports*, 5(5):464–9.

Burge, C. B. and Karlin, S. (1998). Finding the genes in genomic DNA. *Current Opinion in Structural Biology*, 8:346–354.

Buscaino, A., Köcher, T., Kind, J. H., Holz, H., Taipale, M., Wagner, K., Wilm, M., and Akhtar, A. (2003). MOF-regulated acetylation of MSL-3 in the Drosophila dosage compensation complex. *Molecular cell*, 11(5):1265–77.

Butte, A. J., Dzau, V. J., and Glueck, S. B. (2001). Further defining housekeeping, or "maintenance," genes Focus on "A compendium of gene expression in normal human tissues". *Physiological genomics*, 7(2):95–6.

Carrozza, M. J., Li, B., Florens, L., Suganuma, T., Swanson, S. K., Lee, K. K., Shia, W.-J., Anderson, S., Yates, J., Washburn, M. P., and Workman, J. L. (2005). Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. *Cell*, 123(4):581–92.

Chengyin (2008). Tsga: an r package for tissue specific genes analysis. [Online; accessed December 2010].

Core, L. J. and Lis, J. T. (2008). Transcription regulation through promoter-proximal pausing of RNA polymerase II. *Science*, 319(5871):1791–2.

Core, L. J., Waterfall, J., and Lis, J. T. (2008). Nascent rna sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, 322(5909):1845–1848.

Darnell, J. E. (2002). Transcription factors as targets for cancer therapy. *Nature reviews. Cancer*, 2(10):740–9.

DeRisi, J., Penland, L., Brown, P., Bittner, M., Meltzer, P., Ray, M., Chen, Y., Su, Y., and JM, T. (1996). Use of a cdna microarray to analyse gene expression patterns in human cancer. *Nat Genet.*, 14(4):457–60.

Dezso, Z., Nikolsky, Y., Sviridov, E., Shi, W., Serebriyskaya, T., Dosymbekov, D., Bugrim, A., Rakhmatulin, E., Brennan, R. J., Guryanov, A., Li, K., Blake, J., Samaha, R. R., and Nikolskaya, T. (2008). A comprehensive functional analysis of tissue specificity of human gene expression. *BMC biology*, 6:49.

Down, T. a. and Hubbard, T. J. P. (2005). NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence. *Nucleic acids research*, 33(5):1445–53.

Driever, W. and Nüsslein-Volhard, C. (1989). The bicoid protein is a positive regulator of hunchback transcription in the early drosophila embryo. *Nature*, 337(6203):138–43.

Eisenberg, E. and Levanon, E. (2003). Human housekeeping genes are compact. *Trends in Genetics*, 19(7):362–365.

Engelkamp, D. and van Heyningen V. (1996). Transcription factors in disease. *Curr Opin Genet Dev*, 6(3):334–42.

Ernst, J. and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature biotechnology*, 28(8):817–838.

Filion, G. J., van Bemmel, J. G., Braunschweig, U., Talhout, W., Kind, J., Ward, L. D., Brugman, W., de Castro, I. J., Kerkhoven, R. M., and Bussemaker, H. J. (2010). Systematic Protein Location Mapping Reveals Five Principal Chromatin Types in Drosophila Cells. *Cell*.

Fraley, C. and Raftery, A. E. (1999). Mclust: Software for model-based cluster analysis. *Journal of Classification*, 16:297–306.

Fraley, C. and Raftery, A. E. (2003). Enhanced software for model-based clustering, density estimation, and discriminant analysis: Mclust. *Journal of Classification*, 20:263–286.

Fraley, C. and Raftery, A. E. (2006). mclust version 3 for r: Normal mixture modeling and model-based clustering. *Technical Report 504, University of Washington, Department of Statistics*, 20:263–286.

Freilich, S., Massingham, T., Bhattacharyya, S., Ponsting, H., Lyons, P. A., Freeman, T. C., and Thornton, J. M. (2005). Relationship between the tissue-specificity of mouse gene expression and the evolutionary origin and function of the proteins. *Genome Biology*, 6(7):R56.

Fuda, N. J., Ardehali, M. B., and Lis, J. T. (2009). Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature*, 461(7261):186–92.

Furney, S. J., Higgins, D. G., Ouzounis, C. A., and López-Bigas, N. (2006). Structural and functional properties of genes involved in human cancer. *BMC genomics*, 7:3.

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). affy—analysis of affymetrix genechip data at the probe level. *Bioinformatics*, 20(3):307–315.

Gelbart, M. E. and Kuroda, M. I. (2009). Drosophila dosage compensation: a complex voyage to the X chromosome. *Development (Cambridge, England)*, 136(9):1399–410.

Gelbart, M. E., Larschan, E., Peng, S., Park, P. J., and Kuroda, M. I. (2009). Drosophila MSL complex globally acetylates H4K16 on the male X chromosome for dosage compensation. *Nature structural & molecular biology*, 16(8):825–32.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H., and Zhang, J. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80.

Gerstein, M. B., Lu, Z. J., Van Nostrand, E. L., Cheng, C., Arshinoff, B. I., Liu, T., Yip, K. Y., Robilotto, R., Rechtsteiner, a., Ikegami, K., Alves, P., Chateigner, a., Perry, M., Morris, M., Auerbach, R. K., Feng, X., Leng, J., Vielle, a., Niu, W., Rhrissorrakrai, K., Agarwal, a., Alexander, R. P., Barber, G., Brdlik, C. M., Brennan, J., Brouillet, J. J., Carr, a., Cheung, M.-S., Clawson, H., Contrino, S., Dannenberg, L. O., Dernburg, a. F., Desai, a., Dick, L., Dose, a. C., Du, J., Egelhofer, T., Ercan, S., Euskirchen, G., Ewing, B., Feingold, E. a., Gassmann, R., Good, P. J., Green, P., Gullier, F., Gutwein, M., Guyer, M. S., Habegger, L., Han,

T., Henikoff, J. G., Henz, S. R., Hinrichs, a., Holster, H., Hyman, T., Iniguez, a. L., Janette, J., Jensen, M., Kato, M., Kent, W. J., Kephart, E., Khivansara, V., Khurana, E., Kim, J. K., Kolasinska-Zwierz, P., Lai, E. C., Latorre, I., Leahey, a., Lewis, S., Lloyd, P., Lochovsky, L., Lowdon, R. F., Lubling, Y., Lyne, R., MacCoss, M., Mackowiak, S. D., Mangone, M., McKay, S., Mecenas, D., Merrihew, G., Miller, D. M., Muroyama, a., Murray, J. I., Ooi, S.-L., Pham, H., Phippen, T., Preston, E. a., Rajewsky, N., Ratsch, G., Rosenbaum, H., Rozowsky, J., Rutherford, K., Ruzanov, P., Sarov, M., Sasidharan, R., Sboner, a., Scheid, P., Segal, E., Shin, H., Shou, C., Slack, F. J., Slightam, C., Smith, R., Spencer, W. C., Stinson, E. O., Taing, S., Takasaki, T., Vafeados, D., Voronina, K., Wang, G., Washington, N. L., Whittle, C. M., Wu, B., Yan, K.-K., Zeller, G., Zha, Z., Zhong, M., Zhou, X., Ahringer, J., Strome, S., Gunsalus, K. C., Micklem, G., Liu, X. S., Reinke, V., Kim, S. K., Hillier, L. W., Henikoff, S., Piano, F., Snyder, M., Stein, L., Lieb, J. D., and Waterston, R. H. (2010). Integrative Analysis of the Caenorhabditis elegans Genome by the modENCODE Project. *Science*, 330(6012):1775–1787.

Gilchrist, D. A., Nechaev, S., Lee, C., Ghosh, S. K. B., Collins, J. B., Li, L., Gilmour, D. S., and Adelman, K. (2008). Nelf-mediated stalling of pol ii can enhance gene expression by blocking promoter-proximal nucleosome assembly. *Genes & Development*, 20(14):1921–1933.

Gilfillan, G. D., Straub, T., de Wit, E., Greil, F., Lamm, R., van Steensel, B., and Becker, P. B. (2006). Chromosome-wide gene-specific targeting of the Drosophila dosage compensation complex. *Genes & development*, 20(7):858–70.

Gilmour, D. S. (2009). Promoter proximal pausing on genes in metazoans. *Chromosoma*, 118(1):1–10.

Gilmour, D. S. and Lis, J. T. (1986). Rna polymerase ii interacts with the promoter

region of the non induced hsp7o gene in drosophila melanogaster cells. *Nature*, 6(11):3984–3989.

Glozak, M. A., Sengupta, N., Zhang, X., and Seto, E. (2005). Acetylation and deacetylation of non-histone proteins. *Gene*, 363:15–23.

Goffeau, A., Barrell, B., Bussey, H., Davis, R., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J., Jacq, C., Johnston, M., Louis, E., Mewes, H., Murakami, Y., Philippsen, P., Tettelin, H., and Oliver, S. (1996). Life with 6000 genes. *Science*, 274(5287):546–567.

Greco, D., Somervuo, P., Di Lieto, A., Raitila, T., Nitsch, L., Castrén, E., and Auvinen, P. (2008). Physiology, pathology and relatedness of human tissues from gene expression meta-analysis. *PloS one*, 3(4):e1880.

Griffiths, D. S., Li, J., Dawson, M. a., Trotter, M. W. B., Cheng, Y.-H., Smith, A. M., Mansfield, W., Liu, P., Kouzarides, T., Nichols, J., Bannister, A. J., Green, A. R., and Göttgens, B. (2010). LIF-independent JAK signalling to chromatin in embryonic stem cells uncovered from an adult stem cell disease. *Nature cell biology*, 13(1):13–21.

Gu, W., Szauter, P., and Lucchesi, J. C. (1998). Targeting of MOF, a putative histone acetyl transferase, to the X chromosome of Drosophila melanogaster. *Developmental genetics*, 22(1):56–64.

Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R., and Young, R. A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, 130(1):77–88.

Hallacli, E. and Akhtar, A. (2009). X chromosomal regulation in flies: when less is more. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, 17(5):603–19.

Hamada, F. N., Park, P. J., Gordadze, P. R., and Kuroda, M. I. (2005). Global regulation of X chromosomal genes by the MSL complex in Drosophila melanogaster. *Genes & Development*, pages 2289–2294.

He, H. H., Meyer, C. A., Shin, H., Bailey, S. T., Wei, G., Wang, Q., Zhang, Y., Xu, K., Ni, M., Lupien, M., Mieczkowski, P., Lieb, J. D., Zhao, K., Brown, M., and Liu, X. S. (2010). Nucleosome dynamics define transcriptional enhancers. *Nature genetics*, 42(4):343–7.

Hebenstreit, D., Gu, M., Haider, S., Turner, D. J., Liò, P., and Teichmann, S. A. (2010). EpiChIP: gene-by-gene quantification of epigenetic modification levels. *Nucleic acids research*, 3(0):1–10.

Hilfiker, A., Hilfiker-Kleiner, D., Pannuti, A., and Lucchesi, J. C. (1997). mof, a putative acetyl transferase gene related to the Tip60 and MOZ human genes and to the SAS genes of yeast, is required for dosage compensation in Drosophila. *The EMBO journal*, 16(8):2054–60.

Hirose, Y. and Ohkuma, Y. (2007). Phosphorylation of the C-terminal domain of RNA polymerase II plays central roles in the integrated events of eucaryotic gene expression. *Journal of biochemistry*, 141(5):601–8.

Hori, R. and Carey, M. (1994). The role of activators in assembly of RNA polymerase II transcription complexes. *Current opinion in genetics & development*, 4(2):236–44.

Hubbard, T. J. P., Aken, B. L., Ayling, S., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Clarke, L., Coates, G., Fairley, S., Fitzgerald, S., Fernandez-Banet, J., Gordon, L., Graf, S., Haider, S., Hammond, M., Holland, R., Howe, K., Jenkinson, A., Johnson, N., Kahari, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K.,

Meidl, P., Overduin, B., Parker, A., Pritchard, B., Rios, D., Schuster, M., Slater, G., Smedley, D., Spooner, W., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wilder, S., Zadissa, A., Birney, E., Cunningham, F., Curwen, V., Durbin, R., Fernandez-Suarez, X. M., Herrero, J., Kasprzyk, A., Proctor, G., Smith, J., Searle, S., and Flicek, P. (2009). Ensembl 2009. *Nucleic acids research*, 37(Database issue):D690–7.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314.

Jänne, P. A., Li, C., Zhao, X., Girard, L., Chen, T.-H., Minna, J., Christiani, D. C., Johnson, B. E., and Meyerson, M. (2004). High-resolution single-nucleotide polymorphism array and clustering analysis of loss of heterozygosity in human lung cancer cell lines. *Oncogene*, 23(15):2716–26.

John, S., Sabo, P. J., Thurman, R. E., Sung, M.-H., Biddie, S. C., Johnson, T. a., Hager, G. L., and Stamatoyannopoulos, J. a. (2011). Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature genetics*, 43(3).

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316(5830):1497–502.

Kadota, K., Nishimura, S.-I., Bono, H., Nakamura, S., Hayashizaki, Y., Okazaki, Y., and Takahashi, K. (2003). Detection of genes with tissue-specific expression patterns using Akaike's information criterion procedure. *Physiological genomics*, 12(3):251–9.

Kadota, K., Ye, J., Nakai, Y., Terada, T., and Shimizu, K. (2006). ROKU: a novel method for identification of tissue-specific genes. *BMC bioinformatics*, 7:294.

Kauffmann, A., Gentleman, R., and Huber, W. (2009). arrayqualitymetrics–a bio-

conductor package for quality assessment of microarray data. *Bioinformatics*, 25(3):415–416.

Kehle, J. (1998). dMi-2, a Hunchback-Interacting Protein That Functions in Polycomb Repression. *Science*, 282(5395):1897–1900.

Kelley, R. L., Meller, V. H., Gordadze, P. R., Roman, G., Davis, R. L., and Kuroda, M. I. (1999). Epigenetic spreading of the Drosophila dosage compensation complex from roX RNA genes into flanking chromatin. *Cell*, 98(4):513–22.

Kharchenko, P. V., Alekseyenko, A. a., Schwartz, Y. B., Minoda, A., Riddle, N. C., Ernst, J., Sabo, P. J., Larschan, E., Gorchakov, A. a., Gu, T., Linder-Basso, D., Plachetka, A., Shanower, G., Tolstorukov, M. Y., Luquette, L. J., Xi, R., Jung, Y. L., Park, R. W., Bishop, E. P., Canfield, T. P., Sandstrom, R., Thurman, R. E., MacAlpine, D. M., Stamatoyannopoulos, J. a., Kellis, M., Elgin, S. C. R., Kuroda, M. I., Pirrotta, V., Karpen, G. H., and Park, P. J. (2010). Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature*.

Kharchenko, P. V., Tolstorukov, M. Y., and Park, P. J. (2008). Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature biotechnology*, 26(12):1351–9.

Kim, J., Daniel, J., Espejo, A., Lake, A., Krishna, M., Xia, L., Zhang, Y., and Bedford, M. T. (2006). Tudor, MBT and chromo domains gauge the degree of lysine methylation. *EMBO reports*, 7(4):397–403.

Kimura, A., Umehara, T., and Horikoshi, M. (2002). Chromosomal gradient of histone acetylation established by Sas2p and Sir2p functions as a shield against gene silencing. *Nature genetics*, 32(3):370–7.

Kind, J., Vaquerizas, J. M., Gebhardt, P., Gentzel, M., Luscombe, N. M., Bertone,

P., and Akhtar, A. (2008). Genome-wide analysis reveals MOF as a key regulator of dosage compensation and gene expression in Drosophila. *Cell*, 133(5):813–28.

Kirschner, M. A., Arriza, J. L., Copeland, N. G., Gilbert, D. J., Jenkins, N. A., Magenis, E., and Amara, S. G. (1994). The mouse and human excitatory amino acid transporter gene (eaat1) maps to mouse chromosome 15 and a region of syntenic homology on human chromosome 5. *Genomics*, 22(3):631–633.

Korbel, J. O., Tirosh-Wagner, T., Urban, A. E., Chen, X.-N., Kasowski, M., Dai, L., Grubert, F., Erdman, C., Gao, M. C., Lange, K., Sobel, E. M., Barlow, G. M., Aylsworth, A. S., Carpenter, N. J., Clark, R. D., Cohen, M. Y., Doran, E., Falik-Zaccai, T., Lewin, S. O., Lott, I. T., McGillivray, B. C., Moeschler, J. B., Pettenati, M. J., Pueschel, S. M., Rao, K. W., Shaffer, L. G., Shohat, M., Van Riper, A. J., Warburton, D., Weissman, S., Gerstein, M. B., Snyder, M., and Korenberg, J. R. (2009). The genetic architecture of Down syndrome phenotypes revealed by high-resolution analysis of human segmental trisomies. *Proceedings of the National Academy of Sciences of the United States of America*, 106(29):12031–6.

Kouzarides, T. (2007). Chromatin modifications and their function. *Cell*, 128(4):693–705.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French,

L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang,

W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., and Szustakowki, J. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3):R25.

Larschan, E., Alekseyenko, A. A., Gortchakov, A. A., Peng, S., Li, B., Yang, P., Workman, J. L., Park, P. J., and Kuroda, M. I. (2007). MSL complex is attracted to genes marked by H3K36 trimethylation using a sequence-independent mechanism. *Molecular cell*, 28(1):121–33.

Lee, C., Li, X., Hechmer, A., Eisen, M., Biggin, M. D., Venters, B. J., Jiang, C., Li, J., Pugh, B. F., and Gilmour, D. S. (2008). NELF and GAGA factor are linked to promoter-proximal pausing at many genes in Drosophila. *Molecular and cellular biology*, 28(10):3290–300.

Legube, G., Mcweeney, S. K., Lercher, M. J., and Akhtar, A. (2006). X-chromosome-wide profiling of MSL-1 distribution and dosage compensation in Drosophila. *Genes & Development*, pages 871–883.

Liu, X., Yu, X., Zack, D. J., Zhu, H., and Qian, J. (2008). TiGER: a database for tissue-specific gene expression and regulation. *BMC bioinformatics*, 9:271.

López-Bigas, N., Blencowe, B. J., and Ouzounis, C. A. (2006). Highly consistent patterns for inherited human diseases at the molecular level. *Bioinformatics (Oxford, England)*, 22(3):269–77.

Lucchesi, J. C., Kelly, W. G., and Panning, B. (2005). Chromatin remodeling in dosage compensation. *Annual review of genetics*, 39:615–51.

Lukk, M., Kapushesky, M., Nikkilä, J., Parkinson, H., Goncalves, A., Huber, W., Ukkonen, E., and Brazma, A. (2010). A global map of human gene expression. *Nature biotechnology*, 28(4):322–4.

Luscombe, N. M., Austin, S. E., Berman, H. M., and Thornton, J. M. (2000). An overview of the structures of protein-DNA complexes. *Genome biology*, 1(1):REVIEWS001.

Malik, H. S. and Henikoff, S. (2003). Phylogenomics of the nucleosome. *Nature structural biology*, 10(11):882–91.

Maston, G. A., Evans, S. K., and Green, M. R. (2006). Transcriptional regulatory elements in the human genome. *Annual review of genomics and human genetics*, 7:29–59.

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

Maxam, A. M. and Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(2):560–4.

Meller, V. H., Gordadze, P. R., Park, Y., Chu, X., Stuckenholz, C., Kelley, R. L., and Kuroda, M. I. (2000). Ordered assembly of roX RNAs into MSL complexes

on the dosage-compensated X chromosome in Drosophila. *Current biology : CB*, 10(3):136–43.

Mendjan, S. and Akhtar, A. (2007). The right dose for every sex. *Chromosoma*, 116(2):95–106.

Mendjan, S., Taipale, M., Kind, J., Holz, H., Gebhardt, P., Schelder, M., Vermeulen, M., Buscaino, A., Duncan, K., Mueller, J., Wilm, M., Stunnenberg, H. G., Saumweber, H., and Akhtar, A. (2006). Nuclear pore components are involved in the transcriptional regulation of dosage compensation in Drosophila. *Molecular cell*, 21(6):811–23.

Merika, M. and Thanos, D. (2001). Enhanceosomes. *Current Opinion in Genetics & Development*, 11(2):205–208.

Morales, V., Straub, T., Neumann, M. F., Mengus, G., Akhtar, A., and Becker, P. B. (2004). Functional integration of the histone acetyltransferase MOF into the dosage compensation complex. *The EMBO journal*, 23(11):2258–68.

Morrison, D. K. (2009). The 14-3-3 proteins: integrators of diverse signaling cues that impact cell fate and cancer development. *Trends in cell biology*, 19(1):16–23.

Mortazavi, A., Williams, B. A., Mccue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):1–8.

Muse, G. W., Gilchrist, D. a., Nechaev, S., Shah, R., Parker, J. S., Grissom, S. F., Zeitlinger, J., and Adelman, K. (2007). RNA polymerase is poised for activation across the genome. *Nature genetics*, 39(12):1507–11.

Nicol, J. W., Helt, G. A., Erwin, E., Blossom, E., Blanchard, S. G., Chervitz, S. A., Harmon, C., and Loraine, A. E. (2009). Genoviz Software Development Kit: Java

tool kit for building genomics visualization applications. *BMC bioinformatics*, 10(20):266.

Panne, D., Maniatis, T., and Harrison, S. (2007). An atomic model of the interferon-beta enhanceosome. *Cell*, 129(6):1111–23.

Park, P. J. (2009). ChIP-seq: advantages and challenges of a maturing technology. *Nature reviews. Genetics*, 10(10):669–80.

Parkinson, H., Kapushesky, M., Kolesnikov, N., Rustici, G., Shojatalab, M., Abeygunawardena, N., Berube, H., Dylag, M., Emam, I., Farne, A., Holloway, E., Lukk, M., Malone, J., Mani, R., Pilicheva, E., Rayner, T. F., Rezwan, F., Sharma, A., Williams, E., Bradley, X. Z., Adamusiak, T., Brandizi, M., Burdett, T., Coulson, R., Krestyaninova, M., Kurnosov, P., Maguire, E., Neogi, S. G., Rocca-Serra, P., Sansone, S.-A., Sklyar, N., Zhao, M., Sarkans, U., and Brazma, A. (2009). ArrayExpress update–from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic acids research*, 37(Database issue):D868–72.

Pascual, M., Jose, M., Castell, V., and Jover, R. (2008). ATF5 Is a Highly Abundant Liver-Enriched Transcription Factor that Cooperates with Constitutive Androstane Receptor in the Transactivation of CYP2B6 : Implications in Hepatic Stress Responses. *Pharmacology*, 36(6):1063–1072.

Peng, Y., Schwarz, E. J., Lazar, M. A., Genin, a., Spinner, N. B., and Taub, R. (1997). Cloning, human chromosomal assignment, and adipose and hepatic expression of the CL-6/INSIG1 gene. *Genomics*, 43(3):278–84.

Pepke, S., Wold, B., and Mortazavi, A. (2009). Computation for chip-seq and rna-seq studies. *Nature Methods*, 6(11 Suppl).

Peterlin, B. M. and Price, D. H. (2006). Controlling the elongation phase of transcription with P-TEFb. *Molecular cell*, 23(3):297–305.

Press, H., York, N., and Nw, A. (1998). Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. *Science*, 282(5396):2012–2018.

Prestel, M., Feller, C., Straub, T., Mitlöhner, H., and Becker, P. B. (2010). The Activation Potential of MOF Is Constrained for Dosage Compensation. *Molecular Cell*, 38(6):815–826.

Price, D. H. (2008). Poised polymerases: on your mark...get set...go! *Molecular cell*, 30(1):7–10.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raja, S. J., Charapitsa, I., Conrad, T., Vaquerizas, J. M., Gebhardt, P., Holz, H., Kadlec, J., Fraterman, S., Luscombe, N. M., and Akhtar, A. (2010). The nonspecific lethal complex is a transcriptional regulator in Drosophila. *Molecular cell*, 38(6):827–41.

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C. H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J. P., Poggio, T., Gerald, W., Loda, M., Lander, E. S., and Golub, T. R. (2001). Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences of the United States of America*, 98(26):15149–54.

Ravasi, T., Suzuki, H., Cannistraci, C. V., Katayama, S., Bajic, V. B., Tan, K., Akalin, A., Schmeier, S., Kanamori-Katayama, M., Bertin, N., Carninci, P., Daub, C. O., Forrest, A. R. R., Gough, J., Grimmond, S., Han, J.-H., Hashimoto, T., Hide, W., Hofmann, O., Kamburov, A., Kaur, M., Kawaji, H., Kubosaki, A., Lassmann, T., van Nimwegen, E., MacPherson, C. R., Ogawa, C., Radovanovic, A., Schwartz, A., Teasdale, R. D., Tegnér, J., Lenhard, B., Teichmann, S. A., Arakawa, T., Ninomiya, N., Murakami, K., Tagami, M., Fukuda, S., Imamura,

K., Kai, C., Ishihara, R., Kitazume, Y., Kawai, J., Hume, D. A., Ideker, T., and Hayashizaki, Y. (2010). An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, 140(5):744–52.

Reimand, J., Kull, M., Peterson, H., Hansen, J., and Vilo, J. (2007). g:Profiler–a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic acids research*, 35(Web Server issue):W193–200.

Risse, G., Jooss, K., Neuberg, M., Brüller, H. J., and Müller, R. (1989). Asymmetrical recognition of the palindromic AP1 binding site (TRE) by Fos protein complexes. *The EMBO journal*, 8(12):3825–32.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1):139–40.

Robinson, P. J. J., An, W., Routh, A., Martino, F., Chapman, L., Roeder, R. G., and Rhodes, D. (2008). 30 nm chromatin fibre decompaction requires both H4-K16 acetylation and linker histone eviction. *Journal of molecular biology*, 381(4):816–25.

Roh, T.-Y., Cuddapah, S., Cui, K., and Zhao, K. (2006). The genomic landscape of histone modifications in human T cells. *Proceedings of the National Academy of Sciences of the United States of America*, 103(43):15782–7.

Rozowsky, J., Euskirchen, G., Auerbach, R. K., Zhang, Z. D., Gibson, T., Bjornson, R., Carriero, N., Snyder, M., and Gerstein, M. B. (2009). PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nature biotechnology*, 27(1):66–75.

Saltzman, A. and Weinmann, R. (1989). Promoter specificity and modulation of rna polymerase ii transcription. *FASEB J.*, 3(6):1723–33.

Sanger, F. and Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology*, 94(3):441–8.

Schug, J., Schuller, W.-P., Kappen, C., Salbaum, J. M., Bucan, M., and Stoeckert, C. J. (2005). Promoter features related to tissue specificity as measured by Shannon entropy. *Genome biology*, 6(4):R33.

Schwaiger, M., Stadler, M. B., Bell, O., Kohler, H., Oakeley, E. J., and Schu, D. (2009). Chromatin state marks cell-type- and gender-specific replication of the Drosophila genome. *Genes & Development*, pages 589–601.

Schwartz, Y. B., Kahn, T. G., Nix, D. A., Li, X.-Y., Bourgon, R., Biggin, M., and Pirrotta, V. (2006). Genome-wide analysis of Polycomb targets in Drosophila melanogaster. *Nature genetics*, 38(6):700–5.

Shaulian, E. and Karin, M. (2002). AP-1 as a regulator of cell life and death. *Nature cell biology*, 4(5):E131–6.

Shendure, J. and Ji, H. (2008). Next-generation DNA sequencing. *Nature biotechnology*, 26(10):1135–45.

Shin, H., Liu, T., Manrai, A. K., and Liu, X. S. (2009). CEAS: cis-regulatory element annotation system. *Bioinformatics (Oxford, England)*, 25(19):2605–6.

Shlien, A. and Malkin, D. (2009). Copy number variations and cancer. *Genome medicine*, 1(6):62.

Shogren-Knaak, M., Ishii, H., Sun, J.-M., Pazin, M. J., Davie, J. R., and Peterson, C. L. (2006). Histone H4-K16 acetylation controls chromatin structure and protein interactions. *Science*, 311(5762):844–7.

Smale, S. T. and Kadonaga, J. T. (2003). The RNA polymerase II core promoter. *Annual review of biochemistry*, 72:449–79.

Smith, E. R., Allis, C. D., and Lucchesi, J. C. (2001). Linking global histone acetylation to the transcription enhancement of X-chromosomal genes in Drosophila males. *The Journal of biological chemistry*, 276(34):31483–6.

Smith, E. R., Pannuti, a., Gu, W., Steurnagel, a., Cook, R. G., Allis, C. D., and Lucchesi, J. C. (2000). The drosophila MSL complex acetylates histone H4 at lysine 16, a chromatin modification linked to dosage compensation. *Molecular and cellular biology*, 20(1):312–8.

Smyth, G. K. (2005). Limma: linear models for microarray data. In Gentleman, R., Carey, V., Dudoit, S., and R. Irizarry, W. H., editors, *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, pages 397–420. Springer, New York.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular biology of the cell*, 9(12):3273–97.

Statham, A. L., Strbenac, D., Coolen, M. W., Stirzaker, C., Clark, S. J., and Robinson, M. D. (2010). Repitools: an R package for the analysis of enrichment-based epigenomic data. *Bioinformatics*, 26(13):1662–1663.

Straub, T., Gilfillan, G. D., Maier, V. K., and Becker, P. B. (2005). The Drosophila MSL complex activates the transcription of target genes. *Genes & development*, 19(19):2284–8.

Straub, T., Grimaud, C., Gilfillan, G. D., Mitterweger, A., and Becker, P. B. (2008).

The chromosomal high-affinity binding sites for the Drosophila dosage compensation complex. *PLoS genetics*, 4(12):e1000302.

Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M. P., Walker, J. R., and Hogenesch, J. B. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16):6062–7.

Suka, N., Luo, K., and Grunstein, M. (2002). Sir2p and Sas2p opposingly regulate acetylation of yeast histone H4 lysine16 and spreading of heterochromatin. *Nature genetics*, 32(3):378–83.

Sultan, M., Schulz, M. H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., Schmidt, D., O'Keeffe, S., Haas, S., Vingron, M., Lehrach, H., and Yaspo, M.-L. (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, 321(5891):956–60.

Sun, L. E. I., Bartlam, M., Liu, Y., Pang, H. A. I., and Rao, Z. (2005a). Crystal structure of the pyridoxal-5́-phosphate-dependent serine dehydratase from human liver. *Structure*, pages 791–798.

Sun, X., Frierson, H. F., Chen, C., Li, C., Ran, Q., Otto, K. B., Cantarel, B. L., Cantarel, B. M., Vessella, R. L., Gao, A. C., Petros, J., Miura, Y., Simons, J. W., and Dong, J.-T. (2005b). Frequent somatic mutations of the transcription factor ATBF1 in human prostate cancer. *Nature genetics*, 37(4):407–12.

Sural, T. H., Peng, S., Li, B., Workman, J. L., Park, P. J., and Kuroda, I. (2008). The msl3 chromodomain directs a key targeting step for dosage compensation of the drosophila melanogaster x chromosome. *Genetics*, 15(12):1318–1325.

Tekur, S., Pawlak, A., Guellaen, G., and Hecht, N. B. (1999). Contrin, the humam homologue of a germ-cell y-box-binding protein: Cloning, expression and chromosomal localization. *Journal of Andrology*, 20(1):135–144.

The Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055):69–87.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, 98(9):5116–21.

Vakoc, C. R., Mandat, S. A., Olenchock, B. A., and Blobel, G. A. (2005). Histone H3 lysine 9 methylation and HP1gamma are associated with transcription elongation through mammalian chromatin. *Molecular cell*, 19(3):381–91.

van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A. M., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H., and Bernards, R. (2002). A gene-expression signature as a predictor of survival in breast cancer. *The New England journal of medicine*, 347(25):1999–2009.

van Steensel, B. (2011). Chromatin: constructing the big picture. *The EMBO Journal*, 30(10):1885–1895.

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. (2002). Gene expression pro®ling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536.

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., and Luscombe, N. M. (2009). A census of human transcription factors: function, expression and evolution. *Nature reviews. Genetics*, 10(4):252–63.

Vettese-Dadey, M., Grant, P. A., Hebbes, T. R., Crane- Robinson, C., Allis, C. D., and Workman, J. L. (1996). Acetylation of histone H4 plays a primary role in enhancing transcription factor binding to nucleosomal DNA in vitro. *The EMBO journal*, 15(10):2508–18.

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., and Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–6.

Wang, X., Lee, C., Gilmour, D. S., and Gergen, J. P. (2007). Transcription elongation controls cell fate specification in the Drosophila embryo. *Genes & development*, 21(9):1031–6.

Warren, P., Taylor, D., Martini, P. G. V., Jackson, J., and Bienkowska, J. (2007). PANP - a New Method of Gene Detection on Oligonucleotide Expression Arrays. *2007 IEEE 7th International Symposium on BioInformatics and BioEngineering*, pages 108–115.

Warrington, J. A., Nair, A., Mahadevappa, M., and Tsyganskaya, M. (2000). Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. *Physiological genomics*, 2(3):143–7.

Wasserman, W. W. and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, 5(4):276–287.

Wathelet, M. G., Lin, C. H., Parekh, B. S., Ronco, L. V., Howley, P. M., and Maniatis, T. (1998). Virus infection induces the assembly of coordinately activated

transcription factors on the IFN-beta enhancer in vivo. *Molecular cell*, 1(4):507–18.

Weintraub, H., Tapscott, Stephen J.and Davis, R. L., Thayer, Mathew J.and Adam, M. A., Lassar, A. B., and Miller, A. D. (1989). Activation of muscle-specific genes in pigment, nerve, fat, liver, and fibroblast cell lines by forced expression of myod. *Proceedings of the National Academy of Sciences of the United States of America*, 86(14):5434–8.

Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661–678.

Wu, L.-C. (2002). Zas: C2h2 zinc finger proteins involved in growth and development. *Gene Expression*, 10(4):137–152.

Wu, Z., Irizarry, R. A., Gentleman, R., Martinez-Murillo, F., and Spencer, F. (2004). Gcrma a model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association*, 99:909.

Yamada, T., Komoto, J., Kasuya, T., Takata, Y., Ogawa, H., Mori, H., and Takusagawa, F. (2008). A catalytic mechanism that explains a low catalytic activity of serine dehydratase like-1 from human cancer cells: crystal structure and site-directed mutagenesis studies. *Biochimica et biophysica acta*, 1780(5):809–18.

Yamamoto, Y., Kawamoto, T., and Negishi, M. (2003). The role of the nuclear receptor CAR as a coordinate regulator of hepatic gene expression in defense against chemical toxicity. *Archives of biochemistry and biophysics*, 409(1):207–11.

Zeitlinger, J., Stark, A., Kellis, M., Hong, J.-W., Nechaev, S., Adelman, K., Levine, M., and Young, R. A. (2007). RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. *Nature genetics*, 39(12):1512–6.

Zhang, F., Gu, W., Hurles, M. E., and Lupski, J. R. (2009). Copy number variation in human health, disease, and evolution. *Annual review of genomics and human genetics*, 10:451–81.

Zhang, S. (2007). A comprehensive evaluation of SAM, the SAM R-package and a simple modification to improve its performance. *BMC bioinformatics*, 8:230.

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

Zhang, Y. and Oliver, B. (2007). Dosage compensation goes global. *Current opinion in genetics & development*, 17(2):113–20.

A

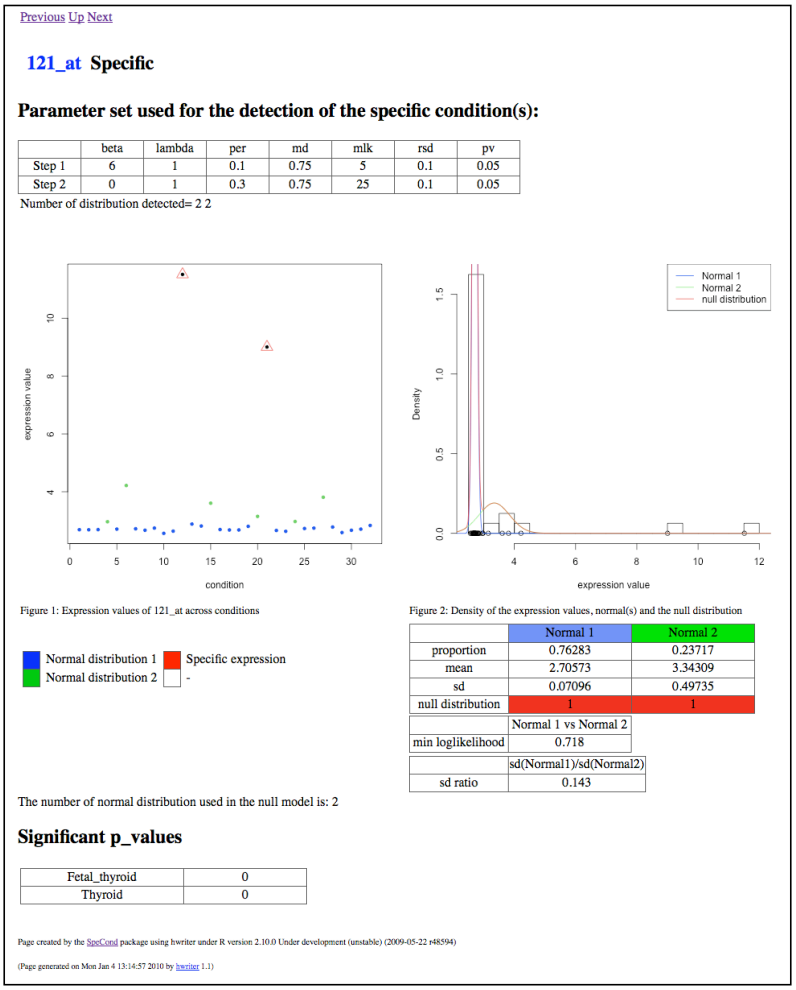
Appendix for Chapter 2

A.1 Table

Tissue	# genes specific		
	up	**down**	**total**
Whole brain	511	4	515
Whole blood	440	9	449
Testis	436	1	437
Fetal brain	406	4	410
Placenta	354	4	358
Liver	287	4	291
Skeletal muscle	279	16	295
Lung	278	1	279
Spinal cord	266	0	266
Fetal liver	261	0	261
Fetal lung	258	1	259
Thymus	249	0	249
Thyroid	244	1	245
Prostate	236	2	238
Smooth muscle	229	4	233
Heart	228	5	233
Bone marrow	227	0	227
Kidney	208	0	208
Uterus	199	1	200
Lymph node	194	1	195
Tonsil	179	0	179
Appendix	171	4	175
Pituitary	165	1	166
Trachea	162	0	162
Tongue	160	0	160
Pancreas	148	0	148
Skin	136	2	138
Adrenal gland	120	0	120
Fetal thyroid	119	1	120
Salivary gland	115	1	116
Adrenal cortex	85	0	85
Ovary	76	0	76

Table A.1: Numbers of tissue specific genes. Numbers of specific genes for 32 human tissues. The detail of the up- and down- and total numbers of specific genes are given.

A.2 Figures



Previous Up Next

121_at Specific

Parameter set used for the detection of the specific condition(s):

	beta	lambda	per	md	mlk	rsd	pv
Step 1	6	1	0.1	0.75	5	0.1	0.05
Step 2	0	1	0.3	0.75	25	0.1	0.05

Number of distribution detected= 2 2

Figure 1: Expression values of 121_at across conditions

Figure 2: Density of the expression values, normal(s) and the null distribution

Normal distribution 1
Normal distribution 2
Specific expression
-

	Normal 1	Normal 2
proportion	0.76283	0.23717
mean	2.70573	3.34309
sd	0.07096	0.49735
null distribution	1	1

	Normal 1 vs Normal 2
min loglikelihood	0.718

	sd(Normal1)/sd(Normal2)
sd ratio	0.143

The number of normal distribution used in the null model is: 2

Significant p_values

Fetal_thyroid	0
Thyroid	0

Page created by the SpeCond package using hwriter under R version 2.10.0 Under development (unstable) (2009-05-22 r48594)

(Page generated on Mon Jan 4 13:14:57 2010 by hwriter 1.1)

Figure A.1: Individual SpeCond specific HTML page; output for a specific probe set. Example of the 121_at probe set detected as specific in 2 tissues. The HTML displays the probe set (or gene) name and a set of tables and figures: the parameters used in the analysis (top table), the expression profile (Figure 1), and the density curves of the mixture model fitting the expression values (Normals 1, 2 and 3, in blue, green and yellow respectively), as well as the null distribution (red) (Figure 2). The parameters of each normal distribution as well as the SpeCond parameter values are presented in the table below the Figure 2. Finally, the tissues in which the gene is detected as specific with their corresponding adjusted p-value are presented in the bottom table.

B

Appendix for Chapter 4

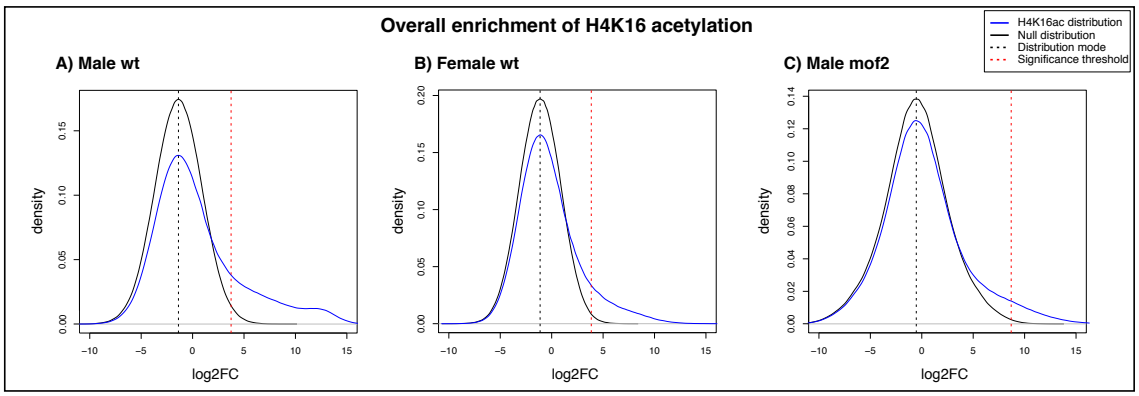


Figure B.1: The overall enrichment of H4K16ac in male, female wt and *mof2* mutant. The density plots show the smoothed density of log2FC values (H4K16ac IP over H4 IP; blue line) with corresponding null distributions (black line) and significance thresholds (red dotted line) for H4K16ac ChIP-seq from male (A), female (B) wild-type and *mof2* mutant salivary glands. The mode of the empirical distribution is indicated as black dotted line.

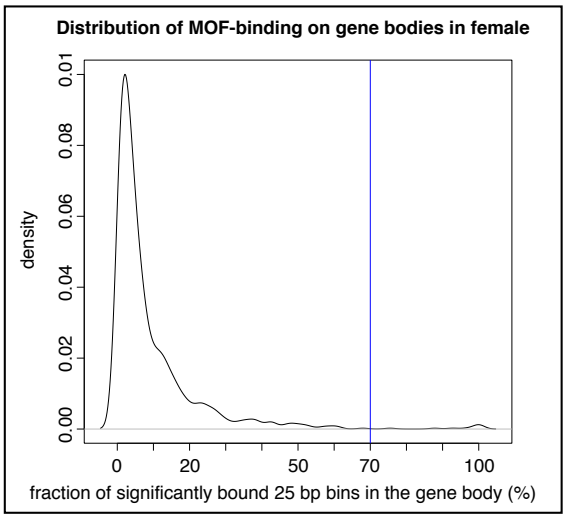


Figure B.2: The distribution of MOF-binding on genes bodies in female flies. The density plots depicts the relative number of gene bodies that show a given percentage of significantly bound 25 bp bins (for this analysis only the exonic regions were considered and the first 500 bp from the TSS were discarded). The blue line indicates the threshold of 70% bound bins to define fully bound genes.

		Sample name	**# reads**	**# reads mapped**	**% reads mapped**	**coverage**
36 bp read		**Male samples:**				
	1	MOF male wt	9,945,648	8,915,892	89.65	1.90
	2	input male wt	7,143,508	6,298,158	88.17	1.34
	3	input male mof2	9,934,408	8,641,131	86.98	1.84
	4	H4 male wt	13,564,190	11,892,074	87.67	2.54
	5	H4K16 male wt	12,567,533	11,038,368	87.83	2.36
	6	H4 male mof2	18,391,608	15,774,286	85.77	3.37
	7	H4K16 male mof2	18,033,157	14,565,538	80.77	3.11
		Female samples:				
	8	MOF female wt	9,935,297	7,008,690	70.54	1.50
	9	Input female wt	15,784,565	13,590,676	86.10	2.90
	10	H4 female wt	12,727,549	11,094,654	87.17	2.37
	11	H4K16 female wt	10,476,721	8,987,154	85.78	1.92
76 bp read		**Male samples:**				
	12	polII maleA1 wt	25,556,049	19,270,618	75.41	8.68
	13	input maleA1 wt	25,936,578	20,906,815	80.61	9.42
	14	polII maleA2 wt	24,802,445	20,376,581	82.16	9.18
	15	input maleA2 wt	24,657,849	20,392,877	82.70	9.19
		Female samples:				
	16	polII femaleA1 wt	24,502,071	18,696,145	76.30	8.42
	17	input femaleA1 wt	29,133,963	23,527,559	80.76	10.60
	18	polII femaleA2 wt	26,164,099	17,055,582	65.19	7.68
	19	input femaleA2 wt	25,881,776	21,181,705	81.84	9.54

Table B.1: ChIP-seq samples with numbers and percentages of reads mapped. This table presents the total numbers of reads sequenced from each samples, the numbers and percentages of the reads aligned by Bowtie Langmead et al. (2009) and selected as they were mapped on one of the *D. melanogaster* chromosome and did not contain any Ns, and the genome coverage obtained from these selected mapped reads.

Differential expression analysis results:

a) Male vs female wt	Up in Female		Down in Female (Up in Male)		Total	
	# genes	% genes	# genes	% genes	# genes	% genes
X (2052)	26	1.26%	14	0.68%	40	2.12%
Auto (10663)	91	0.85%	44	0.41%	135	1.09%
All (12715)	117	0.92%	58	0.46%	175	1.37%

b) Male wt vs mof2	Up, repressed genes		Down, activated genes		Total	
	# genes	% genes	# genes	% genes	# genes	% genes
X (2052)	122	5.95%	403	19.64%	525	25.59%
Auto (10663)	1643	15.41%	721	6.76%	2364	22.17%
All (12715)	1765	13.88%	1124	8.84%	2889	22.7%

c) Female wt vs mof2	Up, repressed genes		Down, activated genes		Total	
	# genes	% genes	# genes	% genes	# genes	% genes
X (2052)	219	10.67%	98	4.77%	317	15.44%
Auto (10663)	1097	10.28%	522	4.89%	1619	15.17%
All (12715)	1316	10.35%	620	4.87%	1936	15.22%

Table B.2: Number and percentage of differentially expressed genes in the three microarray analyses. a) male wt vs female wt, b) male wt vs mof2 mutant, c) female wt vs mof2 mutant.

SpeCond vignette

Condition-specific detection with SpeCond

Florence Cavalli

April 14, 2011

Contents

Introduction

This vignette presents the SpeCond package, an R package which performs gene expression data analysis to detect condition-specific genes. Such genes are significantly up- or down-regulated in a small number of different conditions. Conditions can be environmental conditions, tissues, organs or any other sources that you wish to compare in terms of gene expression.
Condition-specific genes are essentially outliers in an expression pattern. In order to detect such outliers, SpeCond fits a mixture of normal distributions to the expression values. In addition to the main function `SpeCond`, this package includes other methods such as `writeSpeCondResult`, `getFullHtmlSpeCondResult` and `getGeneHtmlPage` which produce text files and HTML reports.

1 The detection process in a few words

The main steps of the procedure are as follows: (i) model the expression data with a mixture of normal distribution(s), (ii) identify the null distribution as well as candidate outlier observations, (iii) compute p-values of the expression values using the null distribution and adjust for multiple comparisons using the Benjami and Yekutiely (or a user-defined) method (iv) identify significant condition-specific expression values of the gene using the adjusted p-values.

2 Quick start

2.1 SpeCond function

The function `SpeCond` enables the user to perform the full analysis. This function is called with the following arguments:

- *expressionMatrix*: an ExpressionSet object or a matrix of expression values (in log2); columns are the conditions, rows are genes (or probe sets)
- *param.detection*: a vector containing the parameters for both steps of the procedure, see section "Parameters"
- *multitest.correction.method*: the multitest correction method; the default is "BY", see `p.adjust` for the possible values
- *prefix.file*: a prefix added to the histogram file (if produced). It will be used to link to the result HTML pages generated by other functions using the result object of this function (if no other prefix value is implemented). The default is "A". It is useful to change the prefix when you perform a new analysis with different parameters, as you may want to compare the results
- *print.hist.pv*: a logical (TRUE/FALSE) value indicating whether a histogram of p-values is to be printed; the default is FALSE
- *condition.factor*: this argument can be used if expressionMatrix is an ExpressionSet object; a factor object of length equal to the number of columns (samples) of the ExpressionSet object specifying which sample(s) belong to which condition (condition.factor levels); can be extracted from the phenoData
- *condition.method*: this argument can be used if expressionMatrix is an ExpressionSet object; the method (mean, median or max) to summarise the samples by conditions (defined by the condition.factor vector)

Example:
Note: For this vignette we will work in a temporary directory (using `tempdir()` as below), to limit the size of the source files. However, for your own use you should ignore it, by default the .RData and files are generated and saved in the current directory.

```
> d=tempdir()
> oldir=getwd()
> setwd(d)
```

Loading the library and the example dataset; an ExpressionSet and a matrix expression value:

```
> library(SpeCond)
```

```
by using mclust, invoked on its own or through another package,
you accept the license agreement in the mclust LICENSE file
and at http://www.stat.washington.edu/mclust/license.txt
```

```
> data(expSetSpeCondExample)
> expSetSpeCondExample
```

```
ExpressionSet (storageMode: lockedEnvironment)
assayData: 220 features, 64 samples
  element names: exprs
protocolData: none
phenoData
  sampleNames: S_1 S_2 ... S_64 (64 total)
  varLabels: Tissue Exp
  varMetadata: labelDescription
featureData: none
experimentData: use 'experimentData(object)'
Annotation:
```

```
> class(expSetSpeCondExample)
```

```
[1] "ExpressionSet"
attr(,"package")
[1] "Biobase"
```

```
> data(expressionSpeCondExample)
> Mexp=expressionSpeCondExample
> class(Mexp)
```

```
[1] "matrix"
```

```
> dim(Mexp)
```

```
[1] 220  32
```

Perform the analysis with default parameters:
Using an expression matrix as input:

```
> generalResult=SpeCond(Mexp, param.detection=NULL, multitest.correction.method="BY",
+   prefix.file="E", print.hist.pv=TRUE, fit1=NULL, fit2=NULL, specificOutlierStep1=NULL)
```

```
[1] "Step1"
[1] "Step1, fitting"
[1] "start: get null distributions"
[1] "end: get null distributions"
[1] "start: specific detection from p-values"
[1] "end: specific detection from p-values"
[1] "Step2"
```

```
[1] "Step2, fitting"
[1] "start: get null distributions"
[1] "end: get null distributions"
[1] "start: specific detection from p-values"
[1] "end: specific detection from p-values"

> names(generalResult)

[1] "prefix.file"           "fit1"                   "fit2"
[4] "specificOutliersStep1" "specificResult"

> specificResult=generalResult$specificResult
```

Using an ExpressionSet object as input:
Specifying the *condition.factor* and the *condition.method* arguments to extract correctly the expression values according to the conditions, see `getMatrixFromExpressionSet` for details.

```
> generalResult=SpeCond(expSetSpeCondExample, param.detection=NULL,
+   multitest.correction.method="BY", prefix.file="E", print.hist.pv=TRUE, fit1=NULL,
+   fit2=NULL, specificOutlierStep1=NULL, condition.factor=expSetSpeCondExample$Tissue,
+   condition.method="mean")
```

Retreive the expression matrix used by the analysis with `getMatrixFronExpressionSet` and the same arguments (*condition.factor* and *condition.method*) used in the `SpeCond` function. This is necessary for the visualisation functions.

```
> MexpS=getMatrixFromExpressionSet(expSetSpeCondExample,
+   condition.factor=expSetSpeCondExample$Tissue,condition.method="mean")
```

The **generalResult** and **specificResult** objects generated above are described in the Output section.

2.2 Result visualisation

Two functions `getFullHtmlSpeCondResult` and `getGeneHtmlPage` allow the user to visualise the results.

2.2.1 HTML page displaying the full detection result

A general result HTML page generated by the `getFullHtmlSpeCondResult` function gives an overall view of the condition specific behaviour of genes present in the dataset. This page mainly contains a plot of the number of specific genes by condition, a specific profile heatmap and the result table with the specific genes and their specific condition(s).

The function `getFullHtmlSpeCondResult` is called with the following arguments:

- *SpeCondResult*: the result object of the `SpeCond` functions
- *param.detection*: the parameter matrix used in the analysis (by `SpeCond`)
- *page.name*: the name of the result HTML page. The default is "SpeCond_result"
- *page.title*: the title of the result HTML page. The default is "Condition-specific analysis results"

Further arguments are described in the paragraph "Visualisation, HTML pages".

```
> getFullHtmlSpeCondResult(SpeCondResult=generalResult, param.detection=
+   specificResult$param.detection, page.name="Example_SpeCond_results",
+   page.title="Tissue specific results", sort.condition="all", heatmap.profile=TRUE,
+   heatmap.expression=FALSE, heatmap.unique.profile=FALSE, expressionMatrix=Mexp)

[1] "The following files are created in the directory:"
[1] "/tmp/RtmpJCsezy/E_General_Result"
[1] "E_barplot_nb_tissue_nb_genes.png"
[1] "E_nb_specific_gene_in_condition.png"
[1] "E_profile_heatmap.png" "E_profile_heatmap.pdf"
[1] "E_result_specific_probeset.txt"
```

2.2.2 HTML result pages for each gene

A result HTML page for each gene can be produced by the function `getGeneHtmlPage`. In addition, this function creates an index to allow you to navigate easily between the results pages for each gene. Each gene result page contains the parameter set used, the expression profile plot and its density curve. An additional plot displays the normal density functions from the normal mixture model as well as the null distribution curve determined by SpeCond. Finally, it presents the condition(s) in which the gene is specific with the associated p-value.

The function `getGeneHtmlPage` is called with the following arguments:

- *expressionMatrix*: the matrix of expression values initially used
- *specificResult*: the R object result of the `getSpecificProbeset` function
- *name.index.html*: the name of the HTML index, the default is "index.html"
- *prefix.file*: a prefix added to the generated file(s) and *outdir* directory name to linked to the index file. The default is NULL, the *prefix.file* attribute of *specificResult* object is used
- *outdir*: the name of the directory in which the generated files will be created. The default is "Single_result_pages"
- *gene.html*: a vector of gene names for which you want to create HTML pages, same as the row names of the *expressionMatrix* object, the default is NULL (the values of the *gene.html.ids* argument will be used)
- *gene.html.ids*: a vector of integers corresponding to the row numbers in the *expressionMatrix* object for the genes for which you want to create HTML pages. The default is the first 10 rows (or the number of rows of the expressionMatrix if lower to 10).

Remark: If both *gene.html* and *gene.html.ids* are set to NULL, the gene HTML pages for every gene in *expressionMatrix* will be generated. It is possible to use *gene.html* or *gene.html.ids* to select a set of genes.
For a given set of genes, you may wish to compare the results contained in two or more **specificResult** objects, which were derived using different parameter sets. For this reason, it is useful to change the prefix (by setting the variable **prefix.file**) as well as the name of the index file (by setting the variable **name.index.html**).

```
> genePageInfo=getGeneHtmlPage(Mexp, specificResult, name.index.html=
+   "index_example_SpeCond_Results.html", gene.html.ids=c(1:20))

[1] "The gene html page(s) will be created in the E_Single_result_pages directory"
```

3 Parameters

3.1 Description

SpeCond uses two sets of parameters that can be tuned intuitively by the user. Two parameters, *lambda* and *beta*, are involved in the implementation of the normal mixture model. Additionally, four parameters are used for the determination of the null distribution: the percentage threshold (*per*), the median difference (*md*), the minimum log-likelihood (*mlk*) and the minimum standard deviation ratio (*rsd*).
To be detected as an outlier, a normal component of the mixed distribution must meet several criteria. The difference between its median and that of the principal normal component (distribution clustering most of the expression values) must be greater than *md*, and the percentage of expression values attributed to this normal must be smaller than *per*. Lastly, the component in question and the principal component must be well-separated, such that the minimum of the absolute log-likelihood ratio of the expression values under the two components is larger than *mlk*, or the component in question must be much more spread-out, such that the ratio of their standard deviations is less than *rsd* (see Figure 1). A p-value threshold is set to define a condition as specific for a given gene. Adjusted p-values are used.

- *lambda*: influences the choice of models by affecting the selection of one, two or three normal distributions, thus introducing some weight on the effect of number of parameters to be defined by the clustering model. The default is 1, the model uses the Bayesian Information Criterion (BIC) value taking into account the log-likelihood value
- *beta*: influences the prior applied during the determination of the variance of the normal distributions. It is necessary in the first fitting step to allow the model to capture isolated outliers. It is set by default to 0 in step 2
- *per*, percentage threshold: this is the maximum percentage of conditions in which a gene can be defined as specific. As per increases, more expression values will be detected as outliers, so each gene can have a larger number of specific conditions associated with it.
- *md*, median difference: this is the minimum distance between the median values of two normal components of the mixted distribution, which allows identification of one component as an outlier (i.e. possibly not part of the null distribution). Low median distances are excluded to filter out noise that is common in biological samples
- *mlk*, minimum log-likelihood: allows the identification of clusters of conditions that are well separated from the other(s) in the model. If the minimum log-likelihood of a set of expression values in a normal component exceeds *mlk*, then this component is a possible outlier (i.e. not part of the null distribution)
- *rsd*, minimum of standard deviation ratio: allows the identification of clusters of conditions that are extremely spread out compared to the distribution clustering of most expression values. If the ratio of standard deviation between the the component and the principale normal component is below *rsd*, then this component is a possible outlier (i.e. is not part of the null distribution)
- *pv*, p-value threshold to detect a condition as specific for a given gene

Figure 1 presents the different detection cases. As *mlk* increases and *rsd* decreases, it is less likely that a mixture component will be defined as an outlier.

3.2 How to change the parameter values

The parameter's values are stored in a matrix (*param.detection*). The procedure includes two steps, which are described in the following section. As a consequence each parameter can be changed to

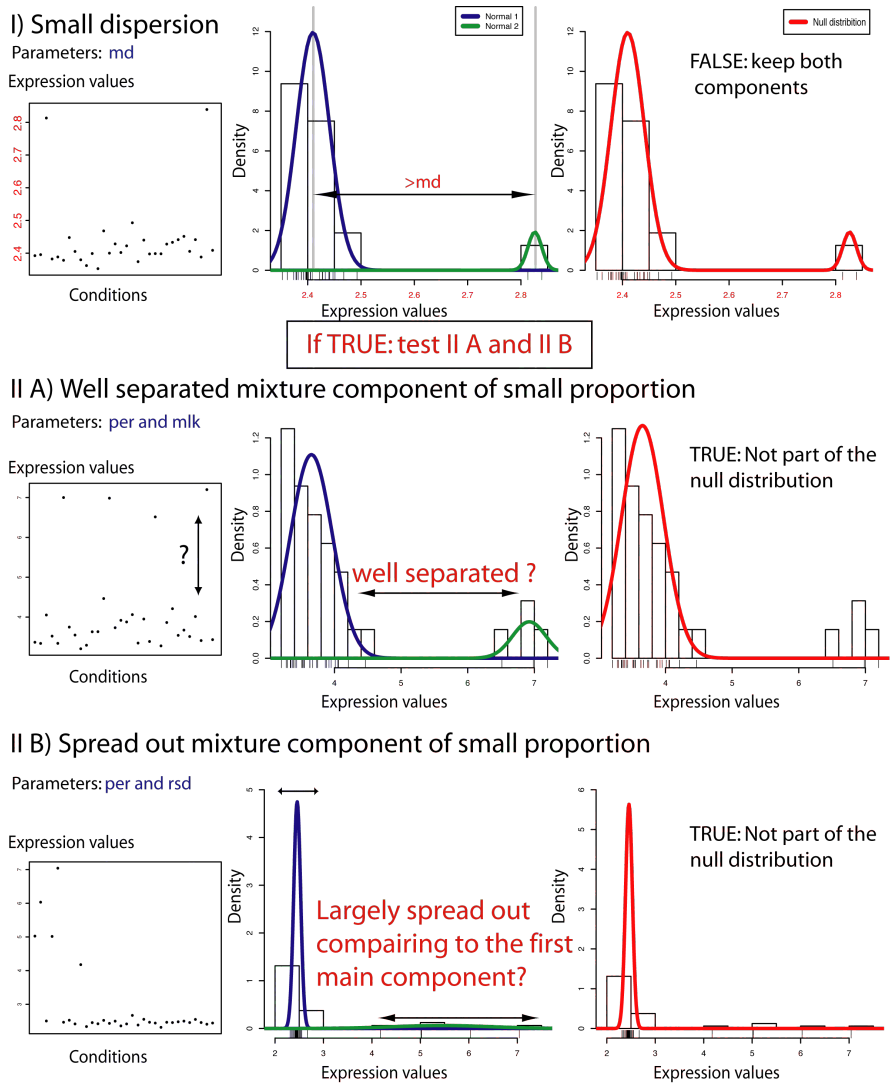


Figure 1: Determination of the null distribution: The three different conditions evaluated in order to consider a normal component as part of the null distribution. (I) The median of the values from each component must have a difference larger than *md*. If this first condition is fulfilled, the procedure tests the following conditions: The normal component will not be part of the null distribution if: (II) The normal component is small and well separated i.e. the minimum of the absolute log-likelihood ratio of the expression values under the two components is larger than *mlk*, (III) The normal component is small and largely spread out i.e. the standard deviation ratio is smaller than *rsd*.

improve the detection. The first row called "Step1" and the second row "Step2" of *param.detection* contain the parameters for the first and the second step of the procedure, respectively.
To obtain the default parameters, use the function `getDefaultParameter`.

```
> param.detection=getDefaultParameter()
> param.detection

       beta lambda per   md mlk rsd   pv
Step 1    6      1 0.1 0.75   5 0.1 0.05
Step 2    0      1 0.3 0.75  25 0.1 0.05
```

The function `createParameterMatrix` enables the user to create a new parameter matrix or to change value(s) from the default or a given parameter matrix. The different arguments allow to change the values in the *param.detection* send as argument. If *param.detection* is not specified (i.e. NULL), the default parameters will be used and then changed according to the other arguments. The different arguments are: *param.detection*, *beta.1*, *beta.2*, *lambda.1*, *lambda.2*, *per.1*, *per.2*, *md.1*, *md.2*, *mlk.1*, *mlk.2*, *rsd.1*, *rsd.2*, *pv.1* and *pv.2*.

```
> param.detection2=createParameterMatrix(mlk.1=10)
> param.detection2

       beta lambda per   md mlk rsd   pv
Step 1    6      1 0.1 0.75  10 0.1 0.05
Step 2    0      1 0.3 0.75  25 0.1 0.05
```

```
> param.detection2B=createParameterMatrix(param.detection=param.detection2,
+   mlk.1=15, rsd.2=0.2)
> param.detection2B

       beta lambda per   md mlk rsd   pv
Step 1    6      1 0.1 0.75  15 0.1 0.05
Step 2    0      1 0.3 0.75  25 0.2 0.05
```

Remark: **We strongly advice the following rules:**

- beta.2=0
- md.1=md.2
- per.1<=per.2
- pv.1=pv.2

3.3 The effect of the parameters on the detections

To give you a sense of the importance and the effect of the parameter values, we have created a simulated dataset. As example, we analyse this dataset with the default parameters then improve the specific detection modifying two parameters. This detailed analysis is present in the "Changing the parameter values to improve the detection results" section.

4 More details

4.1 Advanced usage of SpeCond

The procedure performed by `SpeCond` integrates two steps (presented in Figure 2) that can be processed separately using the functions presented in the following section. Additionally, after running the `SpeCond` function for the first time, it is possible to test other parameters to obtain

the null distribution. In this case, it is not necessary to re-process the fitting step so the results of the first run can be used as an input for the second run. For this purpose, three extra `SpeCond` parameters allow to perform only the latest parts of the procedure (depending of which arguments are still set to NULL).

- *fit1*: the result of the first fitting from the function `fitPrior` or generalResult$fit1, the default is NULL
- *fit2*: the result of the second fitting from the function `fitNoPriorWithExclusion` or generalResult$fit2, the default is NULL
- *specificOutlierStep1*: the result of the first detection step from the function `getSelectiveOutliers` or generalResult$specificOutlierStep1, the default is NULL

Example:
Change the detection parameters for the first step and apply these to the previously computed fitting:

```
> param.detection2=createParameterMatrix(param.detection, mlk.1=10)
> param.detection2

       beta lambda per   md mlk rsd   pv
Step 1    6      1 0.1 0.75  10 0.1 0.05
Step 2    0      1 0.3 0.75  25 0.1 0.05

> generalResult2=SpeCond(Mexp, param.detection=param.detection2,
+   multitest.correction.method="BY", prefix.file="E2",
+   print.hist.pv=TRUE, fit1=generalResult$fit1, fit2=NULL,
+   specificOutlierStep1=NULL)

[1] "Step1"
[1] "start: get null distributions"
[1] "end: get null distributions"
[1] "start: specific detection from p-values"
[1] "end: specific detection from p-values"
[1] "Step2"
[1] "Step2, fitting"
[1] "start: get null distributions"
[1] "end: get null distributions"
[1] "start: specific detection from p-values"
[1] "end: specific detection from p-values"
```

Change the parameters for the second step and apply these to the result of the first step and the second fitting computed previously:

```
> param.detection3=createParameterMatrix(param.detection, rsd.2=0.2, per.2=0.2)
> param.detection3

       beta lambda per   md mlk rsd   pv
Step 1    6      1 0.1 0.75   5 0.1 0.05
Step 2    0      1 0.2 0.75  25 0.2 0.05

> generalResult3=SpeCond(Mexp, param.detection=param.detection3,
+   multitest.correction.method="BY", prefix.file="E3", print.hist.pv=TRUE,
+   fit1=generalResult$fit1, fit2=generalResult$fit2,
+   specificOutlierStep1=generalResult$specificOutliersStep1)
```

```
[1] "Step1"
[1] "Use specificOutlierStep1 as result of step1"
[1] "Step2"
[1] "start: get null distributions"
[1] "end: get null distributions"
[1] "start: specific detection from p-values"
[1] "end: specific detection from p-values"
```

Step 1: Detect "single" outliers
- Fitting with beta.1!=0
  To be able to fit a normal distribution with one or two values as consequence to catch "single" outliers

- Detection: per.1=0.1
  Must be small representing 1 to 3 conditions to be able to detect and test the specificity (using the others paramters) of the normal components containing the "single" outliers

Step 2: Detect the specific-genes
- Fitting with beta.2=0
  The "single" outliers detected in Step1 are ignored

- Detection: per.2=0.3
  Performs the final detection here 30% of the conditions can be detected as specific for a particular gene

Figure 2: Procedure Steps: Short description of the two step procedure indicating key points in each steps

4.2 Stepwise analysis

As the method works with two steps it can be very useful to process step by step, saving the object(s) and enabling to re-run only the last detection function for example with a new set of parameters.
Here, we described the procedure step by step:
Use the function `getMatrixFromExpressionSet` if your dataset is an ExpressionSet to obtain the **Mexp** matrix as presented in the "Quick Start" paragraph.
Step1: get the mixture distributions fitting the expression data with a prior on the variance to catch the single outliers. Then detected the single outliers using the first row ("Step1") of the *param.detection* matrix:

```
> param.detection
```

```
       beta lambda per   md mlk rsd   pv
Step 1    6      1 0.1 0.75   5 0.1 0.05
Step 2    0      1 0.3 0.75  25 0.1 0.05
```

```
> fit1=fitPrior(Mexp, param.detection=param.detection)
```

```
[1] "Step1, fitting"
```

Or specifying only *lambda* and *beta* values:

```
> fit1=fitPrior(Mexp, lambda=param.detection[1,"lambda"], beta=param.detection[1,"beta"])
```

Detect the single outliers and store them in **specificOutlierStep1**.

```
> specificOutlierStep1=getSpecificOutliersStep1(Mexp, fit=fit1$fit1,param.detection,
+   multitest.correction.method="BY", prefix.file="run1_Step1",
+   print.hist.pv=FALSE)

[1] "start: get null distributions"
[1] "end: get null distributions"
[1] "start: specific detection from p-values"
[1] "end: specific detection from p-values"
```

Step2: get the second fitting ignoring the values of outliers detected in step 1, then perform the specific detection using the second row ("Step2") of the param.detection matrix:

```
> fit2=fitNoPriorWithExclusion(Mexp, specificOutlierStep1=specificOutlierStep1,
+   param.detection=param.detection)

[1] "Step2, fitting"
```

Or specifying only *lambda* and *beta* values:

```
> fit2=fitNoPriorWithExclusion(Mexp, specificOutlierStep1=specificOutlierStep1,
+   lambda=param.detection[2,"lambda"], beta=param.detection[2,"beta"])
```

Detect the condition specific for each gene and store them in **specificResult**.

```
> specificResult=getSpecificResult(Mexp, fit=fit2,
+   specificOutlierStep1=specificOutlierStep1, param.detection,
+   multitest.correction.method="BY", prefix.file="run1_Step2",
+   print.hist.pv=FALSE)

[1] "start: get null distributions"
[1] "end: get null distributions"
[1] "start: specific detection from p-values"
[1] "end: specific detection from p-values"
```

5 Output

The SpeCond package can produce three different types of outputs: R objects, text files and HTML pages.

5.1 R Objects

The result of `SpeCond` the **generalResult** object is of class **sp_list**. It is a large list containing all the parameters and results of the analysis. The five attributes of **generalResult** correspond to the output of the functions presented in the Stepwise analysis paragraph. So the object **specifcResult** obtained by the `getSpecificResult` function is a class **sp_list** as well and corresponds to the fifth attributes of the **generalResult** of the `SpeCond` function. The later is a list of seven attributes containing all the parameters and results of the detection.
Remark: For all result objects the order of genes and conditions from the input expression value matrix is preserved.
The **generalResult** attributes are:

- *prefix.file*: the prefix used for this analysis. It will be used by default in the function `getFullHtmlSpeCondResult` and `getGeneHtmlPage`
- *fit1*: a list of genes as first attributes and for each gene a list of three attributes:
  - *G*: number of normal components fitting the data
  - *NorMixParam*: the parameters of each normal component of the mixture distribution fitting the expression values of the gene: proportion, mean and standard deviation
  - *classification*: the normal component of the mixture distribution, to which the expression value is attributed
  - Remark: if *evaluate.lamda.beta* is TRUE in `fitPrior`, the result object will be a list of two arguments:
    - *fit1*: same list as presented above with *G*, *NorMix_param* and *classification* attributes for each gene
    - *G.lambda.beta.effect*: matrix presenting the number of times the values of G (number of normal components for a particular gene) have changed between *lambda*=0 and the *lambda.1* value and between *beta*=0 and the *beta.1* value
- *fit2*: same list as *fit1* described above with attributes: *G_initial*, *G*,*NorMixParam* and *classification* for each gene and an additional attribute *specificOutlierStep1*:
  - *specificOutlierStep1*: the condition(s) for which the expression value of the gene is detected as outlier in the first step of the procedure. If NULL, no expression value has been detected in the first step. The second fitting ignores these expression values
- *specificOutliersStep1*: a list of all genes with the condition(s) (i.e. column id(s)), in which the gene has been detected as specific, NULL if not
- *specificResult*: described below

The **specificResult** object is of class **sp_list**. This list containing 8 attributes

- *prefix.file*: the prefix used for this analysis. It will be used by default in the function `getGeneHtmlPage`
- *param.detection*: the parameters used for the two steps
- *fit*: the fit object of the second step (same as *fit2*)
- *L.specific.result*: full detection results (it will be used by the `getFullHtmlSpeCondResult` function). This list contains 7 attributes:
  - *M.specific.all*: matrix of 0: not selective, 1: selective up-regulated, -1: selective down-regulated; same dimensions as the input matrix of expression values
  - *M.specific*: same as *M.specific.all* but reduced to the specific genes. NULL if no gene has been detected as specific
  - *M.specific.sum.row*: Number of conditions in which the gene is specific (-1 and 1 values)
  - *M.specific.sum.column*: Number of specific genes by conditions
  - *L.pv*: list of all genes with a matrix of conditions and the corresponding p-values (if the gene is specific)
  - *specific*: vector of size the number of genes with only the values "Not specific" or "Specific" according to the specificity of the gene (used by the HTML display)
  - *L.condition.specific.id*: list of the specific genes with a vector of column numbers (condition ids), for which the gene is specific

- *L.null*: a list containing a vector of 1 and 0 representing the null distribution. The length of the vector for each gene corresponds to the number of normal components fitting the gene expression value. The list is sorted as the gene in the input matrix of expression values
- *L.mlk*: a list of vectors containing the minimum log-likelihood computed between normal distributions. NULL if the mixture model of the gene is composed of only one component or if the proportion of all components is superior to the *per.2* parameter
- *L.rsd*: a list of vectors containing the standard deviation ratios computed between normal distributions. NULL if the mixture model of the gene has only one component
- *identic.row.ids*: row number(s) from the initial input matrix, which contain identical values for all conditions. These rows are not considered in the analysis

Example:

```
> names(generalResult)

[1] "prefix.file"           "fit1"                  "fit2"
[4] "specificOutliersStep1" "specificResult"

> specificResult=generalResult$specificResult
> names(specificResult)

[1] "prefix.file"       "fit"               "param.detection"
[4] "L.specific.result" "L.null"            "L.mlk"
[7] "L.rsd"             "identic.row.ids"

> names(specificResult$L.specific.result)

[1] "M.specific.all"          "M.specific"
[3] "M.specific.sum.row"      "M.specific.sum.column"
[5] "L.pv"                    "specific"
[7] "L.condition.specific.id"

> dim(specificResult$L.specific.result$M.specific.all)

[1] 220  32

> dim(specificResult$L.specific.result$M.specific)

[1] 23 32

> specificResult

An object of class "sp_list"
 Only the main values are presented here see show.sp_list() function for
a comprehensive view of this object
$prefix.file
[1] "E"
$param.detection
       beta lambda per   md mlk rsd   pv
Step 1    6      1 0.1 0.75   5 0.1 0.05
Step 2    0      1 0.3 0.75  25 0.1 0.05
$L.specific.result$M.specific
[1] "M.specific"
            Spinal_cord Fetal_brain Adrenal_cortex Pituitary Whole_brain
204501_at             0           0              1         0           0
```

```
204507_s_at           0           0              0         0           0
204508_s_at           0           0              0         0           0
204529_s_at           0           0              0         0           0
204532_x_at           0           0              0         0           0
18 more rows and 27 more columns ...

$L.specific.result$M.specific.sum.row
  200606_at 200607_s_at 200608_s_at   204501_at 204507_s_at
          0           0           0           2           1
215 more elements ...

$L.specific.result$M.specific.sum.column
                        Spinal_cord Fetal_brain Adrenal_cortex Pituitary
M_specific_positive_sum           4           6              1         4
M_specific_negative_sum           0           0              0         0
Both                              4           6              1         4
                        Whole_brain
M_specific_positive_sum           5
M_specific_negative_sum           0
Both                              5
27 more columns ...


Number of genes evaluated: 220
Number of genes specific: 23
Range of the number of conditions for which a gene have been detected as specific: 0 8
Number of genes specific by number of specific conditions
             [,1] [,2] [,3] [,4] [,5] [,6] [,7]
# conditions "1"  "2"  "5"  "6"  "7"  "8"  "sum"
# genes      "9"  "7"  "4"  "1"  "1"  "1"  "23"
```

5.2 Text files

Three functions enable to write the results to text files.

- `writeSpeCondResult`: To write the three following text files:
  - The table of genes detected as specific and in which condition they are specific (0: not specific, 1: specific up-regulated, -1: specific down-regulated). The file name is set by the parameter *file.name.profile*, the default name is "specific_profile.txt".
  - The list of the specific genes. The file name is set by the parameter *file.specific.gene*, the default name is: "list_specific_probeset.txt".
  - The table of the unique specific profiles detected. The file name is set by the parameter *file.name.unique.profile*, the default name is: "specific_unique_profile.txt".
- `writeUniqueProfileSpecificResult`: To write a text file with the unique specific profiles among the conditions. If *full.list.gene* is TRUE, the last column corresponds to the gene's names which have the profile described in the row.
- `writeGeneResult`: To write a text file containing the list of all genes from the input expression values matrix, whether they have been detected as condition specific or not (S/N), for how many conditions in total, in how many conditions as up-regulated, in how many conditions as down-regulated, in which conditions as up-regulated and down-regulated. The argument *gene.names* can be used to select a subset of genes (the default is NULL).

All these functions use the function `getProfile` to obtain the profiles of the specific genes, see documentation and example below

```
> L.specific.result.profile=getProfile(specificResult$L.specific.result$M.specific)
> writeSpeCondResult(specificResult$L.specific.result,file.name.profile=
+   "Example_specific_profile.txt", file.specific.gene="Example_list_specific_gene.txt",
+   file.name.unique.profile="Example_specific_unique_profile.txt")

[1] "write file: Example_specific_profile.txt"
[1] "write file: Example_list_specific_gene.txt"
[1] "write file: Example_specific_unique_profile.txt"

> writeUniqueProfileSpecificResult(L.specific.result=specificResult$L.specific.result,
+   file.name.unique.profile="Example_specific_unique_profile.txt", full.list.gene=FALSE)

[1] "write file: Example_specific_unique_profile.txt"

> writeGeneResult(specificResult$L.specific.result, file.name.result.gene=
+   "Example_gene_gummary_result.txt", gene.names=rownames(Mexp)[1:10])

[1] "write file: Example_gene_gummary_result.txt"
```

5.3 Visualisation, HTML pages

The HTML page functions `getFullHtmlSpeCondResult` and `getGeneHtmlPage` have been presented above. Here we present all the possible arguments of these functions enabling to personalise the result HTML pages.

5.3.1 getFullHtmlSpeCondResult

Some other arguments can be used in the `getFullHtmlSpeCondResult`; notably the arguments *prefix.file* and *outdir* enable to properly save the different result pages.

- *SpeCondResult*: the result object of **sp_list** class result of the `SpeCond` functions
- *L.specific.result*: list of results present in the **specificResult sp_list** class object , see `SpeCond` or `getSpecificResult` functions
- *param.detection*: the parameter matrix used in the analysis (by `SpeCond`)
- *page.name*: the name of the result HTML page. The default is "SpeCond_result"
- *page.title*: the title of the result HTML page. The default is "Condition-specific analysis results"
- *prefix.file*: a prefix added to the generated file(s) and the *outdir* directory name to linked them to the full result HTML page. The default is NULL. It is useful to change the prefix when you create a new result page. As you may want to get results with different parameter sets and plots so using a different *SpeCondResult* or *L.specific.result* objects
- *outdir*: the name of the directory in which the generated files will be created. The default is "General_result"
- *sort.condition*: the way to sort the conditions in the barplot that presents the number of specific genes per condition. Values are "positive", "negative" or "all" determines that the conditions are sorted by the number of specific genes detected as up-regulated, down-regulated or both respectively

- *gene.page.info*: the result of the `getGeneHtmlPage` function. Enables the creation of links between this full result page and the single result pages created by the previous function. The default is "NULL"; no links are created
- *heatmap.profile*: a logical (TRUE/FALSE) whether to print or not a heatmap showing the specific profile of the genes, the default is FALSE
- *heatmap.expression*: a logical (TRUE/FALSE) whether to print or not a heatmap showing the expression of the genes, the default is FALSE
- *heatmap.unique.profile*: a logical (TRUE/FALSE) whether to print or not a heatmap showing the unique specific profile, the default is FALSE
- *expressionMatrix*: Must not be NULL if heatmap.expression=TRUE, must be the same as the input expression matrix, the default is NULL

Remark: One can use either *SpeCondResult* or *L.specific.result* as the *L.specific.result* is included in the *SpeCondResult* object. If *prefix.file* is NULL the *prefix.file* attribute of *SpeCondResult* is used.

5.3.2 getGeneHtmlPage

The function `getGeneHtmlPage` creates one HTML page by genes and the corresponding plots. As a consequence a large amount of files can be generated. To keep track of the result pages linked to different parameter sets, it is important to use the *prefix.file* arguments. The main index HTML page will be created in the current directory whereas all the other gene HTML pages will be created in the *outdir* directory.
Here are all the `getGeneHtmlPage` arguments, more importantly the arguments *prefix.file* and *outdir* allow to configure the classification of your results.

- *expressionMatrix*: the matrix of expression values initially used
- *specificResult*: the R object result of the `getSpecificProbeset` function
- *name.index.html*: the name of the HTML index, the default is "index.html"
- *prefix.file*: a prefix added to the generated file(s) and *outdir* directory name to linked to the index file. The default is NULL, the *prefix.file* attribute of the *specificdResult* is used
- *outdir*: The name of the directory in which the generated files will be created. The default is "Single_result_pages"
- *gene.html*: a vector of gene names, same as the row names of the *expressionMatrix* object, the default is NULL
- *gene.html.ids*: a vector of integers corresponding to the row numbers in the *expressionMatrix* object for the gene for which you want to create HTML pages. The default is the 10 first rows (or the number of row of the expressionMatrix if inferior to 10)

6 Changing the parameter values to improve the detection results

In this section, we will use a simulated dataset to appreciate the effect of the parameters. The principal parameters of the detection are *mlk.2* and *rsd.2*.
The dataset is simulated from three different normal distributions. The default expression values for each probeset is randomly generated from a normal distribution of mean=7 and sd=0.6. The probesets 1 to 100 have specific expression values for the conditions 10, 20 and 30 coming from a

normal distribution of mean=11 and sd=0.5. The probesets 200 to 300 have specific expression values for the conditions 9, 18 and 27 coming from a normale distribution of mean=13 and sd=0.4. We will change the parameter *mlk.2* and *rsd.2* to improve the specific detection (detecting the two types of specific behavior) starting with the default parameters. (The following code is present the *SpeCond.R* file available on the Bioconductor package page.)

Load the simulated dataset:

```
> data(simulatedSpeCondData)
```

Perform the first analysis using the default parameters:

```
> generalResult_S1=SpeCond(simulatedSpeCondData, param.detection=NULL,
+   multitest.correction.method="BY", prefix.file="S1", print.hist.pv=TRUE,
+   fit1=NULL, fit2=NULL, specificOutlierStep1=NULL)
```

```
[1] "Step1"
[1] "Step1, fitting"
[1] "start: get null distributions"
[1] "end: get null distributions"
[1] "start: specific detection from p-values"
[1] "end: specific detection from p-values"
[1] "Step2"
[1] "Step2, fitting"
[1] "start: get null distributions"
[1] "end: get null distributions"
[1] "start: specific detection from p-values"
[1] "end: specific detection from p-values"
```

```
> specificResult_S1=generalResult_S1$specificResult
> getFullHtmlSpeCondResult(L.specific.result=specificResult_S1$L.specific.result,
+   page.name="simulatedSpeCondData_results", page.title="Condition specific results mlk 25
+   default", prefix.file="S1", sort.condition="all")
```

```
[1] "Advice: Implement the param.detection attribute or used the SpeCondResult attribute"
[1] "to present the parameter set used on the S1_simulatedSpeCondData_results.html page."
[1] "The following files are created in the directory:"
[1] "/tmp/RtmpJCsezy/S1_General_Result"
[1] "S1_barplot_nb_tissue_nb_genes.png"
[1] "S1_nb_specific_gene_in_condition.png"
[1] "S1_profile_heatmap.png" "S1_profile_heatmap.pdf"
[1] "S1_result_specific_probeset.txt"
```

```
> genePageInfo_S1=getGeneHtmlPage(simulatedSpeCondData, specificResult_S1,
+   name.index.html="index_simulatedSpeCondData.html", prefix.file="S1",
+   gene.html.ids=c(1:20))
```

```
[1] "The gene html page(s) will be created in the S1_Single_result_pages directory"
```

We open *S1_simulatedSpeCondData_results.html* and *S1_index_simulatedSpeCondData.html* files present in the current directory to observe the results.
Looking at the first results, only few probeset are detected as specific. If you look at the gene result pages you can observe that the *mlk* value of the Step2 in particular is too high as the Normal 2 is still part of the null distribution whereas it can be consider as separated from the Normal 1. As a consequence we change the *mlk.2* parameter to 10. Additionally we are using here the result of the fitting from the first detection as described in the Stepwise analysis paragraph.

Change the parameter:

```
> param.detection10=createParameterMatrix(param.detection=param.detection, mlk.2=10)
> param.detection10

       beta lambda per   md mlk rsd   pv
Step 1    6      1 0.1 0.75   5 0.1 0.05
Step 2    0      1 0.3 0.75  10 0.1 0.05
```

Perform the analysis changing the prefix value:

```
> generalResult_S2=SpeCond(simulatedSpeCondData, param.detection=param.detection10,
+   multitest.correction.method="BY", prefix.file="S2", print.hist.pv=TRUE,
+   fit1=generalResult_S1$fit1, fit2=generalResult_S1$fit2,
+   specificOutlierStep1=generalResult_S1$specificOutliersStep1)

[1] "Step1"
[1] "Use specificOutlierStep1 as result of step1"
[1] "Step2"
[1] "start: get null distributions"
[1] "end: get null distributions"
[1] "start: specific detection from p-values"
[1] "end: specific detection from p-values"

> specificResult_S2=generalResult_S2$specificResult
> genePageInfo_S2=getGeneHtmlPage(simulatedSpeCondData, specificResult_S2,
+   name.index.html="index_simulatedSpeCondData.html", prefix.file="S2",
+   gene.html.ids=c(1:20,200:250))

[1] "The gene html page(s) will be created in the S2_Single_result_pages directory"

> getFullHtmlSpeCondResult(L.specific.result=specificResult_S2$L.specific.result,
+   param.detection=param.detection10, page.name="simulatedSpeCondData_results",
+   page.title="Condition specific results mlk 10",prefix.file="S2", sort.condition="all",
+   gene.page.info=genePageInfo_S2)

[1] "The following files are created in the directory:"
[1] "/tmp/RtmpJCsezy/S2_General_Result"
[1] "S2_barplot_nb_tissue_nb_genes.png"
[1] "S2_nb_specific_gene_in_condition.png"
[1] "S2_profile_heatmap.png" "S2_profile_heatmap.pdf"
[1] "S2_result_specific_probeset.txt"
```

Looking at *S2_simulatedSpeCondData_results.html* and *S2_index_simulatedSpeCondData.html* files we observe that the detection improves but we are still missing some specific probesets with ids between 1 and 100 and more importantly we do not yet detect several specific probe with ids between 200 and 300 for which their specific values are more spread than the core of the expression values. To detect them we decrease *mlk.2* and increase *rsd.2* to 8 and to 0.3 respectively.

Change the parameter:

```
> param.detection8_0.3=createParameterMatrix(param.detection=param.detection,
+   mlk.2=8, rsd.2=0.3)
> param.detection8_0.3

       beta lambda per   md mlk rsd   pv
Step 1    6      1 0.1 0.75   5 0.1 0.05
Step 2    0      1 0.3 0.75   8 0.3 0.05
```

Perform the analysis changing the prefix value:

```
> generalResult_S3=SpeCond(simulatedSpeCondData, param.detection=param.detection8_0.3,
+   multitest.correction.method="BY", prefix.file="S3", print.hist.pv=TRUE,
+   fit1=generalResult_S1$fit1, fit2=generalResult_S1$fit2,
+   specificOutlierStep1=generalResult_S1$specificOutliersStep1)

[1] "Step1"
[1] "Use specificOutlierStep1 as result of step1"
[1] "Step2"
[1] "start: get null distributions"
[1] "end: get null distributions"
[1] "start: specific detection from p-values"
[1] "end: specific detection from p-values"

> specificResult_S3=generalResult_S3$specificResult
> genePageInfo_S3=getGeneHtmlPage(simulatedSpeCondData, specificResult_S3,
+   name.index.html="index_simulatedSpeCondData.html", prefix.file="S3",
+   gene.html.ids=c(1:20,195:310))

[1] "The gene html page(s) will be created in the S3_Single_result_pages directory"

> getFullHtmlSpeCondResult(L.specific.result=specificResult_S3$L.specific.result,
+   param.detection=param.detection8_0.3, page.name="simulatedSpeCondData_results",
+   page.title="Condition specific results mlk 8, rsd 0.3", prefix.file="S3",
+   sort.condition="all",gene.page.info=genePageInfo_S3)

[1] "The following files are created in the directory:"
[1] "/tmp/RtmpJCsezy/S3_General_Result"
[1] "S3_barplot_nb_tissue_nb_genes.png"
[1] "S3_nb_specific_gene_in_condition.png"
[1] "S3_profile_heatmap.png" "S3_profile_heatmap.pdf"
[1] "S3_result_specific_probeset.txt"
```

Looking at *S3_simulatedSpeCondData_results.html* and *S3_index_simulatedSpeCondData.html* files we observe that the detection has largely improved. Change the *gene.html.ids* or *gene.html* values to generate different HTML pages and keep improving the detection results.

7 Advice

The fitting of the normal component using mclust can be done only once, in each of the steps of the procedure, as a consequence it is beneficial to save the fitting results as shown below.

```
> save(fit1,file="fit1.RData")
> ## load("fit1.RData")
```

The parameter effect can be evaluated separately for the two steps. Saving the outliers detected by the first steps allows to test only the effect of the second set of parameters.

```
> save(specificOutlierStep1,file="specificOutlierStep1.RData")
> ## load("specificStep1.RData")
> save(fit2,file="fit2.RData")
> ## load("fit2.Rdata")
> save(specificResult,file="specificResult.RData")
```

Session Info

```
> toLatex(sessionInfo())
```

- R version 2.13.0 (2011-04-13), x86_64-unknown-linux-gnu
- Locale: LC_CTYPE=en_US.UTF-8, LC_NUMERIC=C, LC_TIME=en_US.UTF-8, LC_COLLATE=C, LC_MONETARY=C, LC_MESSAGES=en_US.UTF-8, LC_PAPER=en_US.UTF-8, LC_NAME=C, LC_ADDRESS=C, LC_TELEPHONE=C, LC_MEASUREMENT=en_US.UTF-8, LC_IDENTIFICATION=C
- Base packages: base, datasets, grDevices, graphics, methods, stats, utils
- Other packages: Biobase 2.12.0, RColorBrewer 1.0-2, SpeCond 1.6.0, fields 6.3, hwriter 1.3, mclust 3.4.8, spam 0.23-0
- Loaded via a namespace (and not attached): tools 2.13.0

8 References

Florence MG Cavalli, Juan-Manuel Vaquerizas, Richard Bourgon, Nicholas M Luscombe (in preparation)

C. Fraley and A. E. Raftery, Model-based clustering, discriminant analysis, and density estimation, *Journal of the American Statistical Association, Vol. 97, pages 611-631 (2002).*

C. Fraley and A. E. Raftery, MCLUST Version 3 for R: Normal Mixture Modeling and Model-based Clustering, Technical Report No. 504, Department of Statistics, University of Washington, September 2006.