Analysis and modeling of the ecdysone response in Drosophila melanogaster

Roberto Cortini

München 2018

Dissertation zur Erlangung des Doktorgrades
der Fakultät für Chemie und Pharmazie
der Ludwig–Maximilians–Universität
München

Analysis and modeling of the ecdysone response in Drosophila melanogaster

von
Roberto Cortini
aus Figline Valdarno, Italien

München

Erklärung:

Diese Dissertation wurde im Sinne von §7 der Promotionsordnung vom 28. November 2011 von Prof. Dr. Roland Beckmann betreut.

Eidesstattliche Versicherung:

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, den 9.10.2018

Roberto Cortini

Dissertation eingereicht am 9.10.2018
Erstgutachter: Prof. Dr. Roland Beckmann
Zweitgutachter: Dr. Stefan Canzar
Tag der mündlichen Prüfung: 12.11.2018

Contents

Acknowledgments

I would like to start my ‘thanks‘ by thanking my supervisor Ulrike Gaul. She trusted me and gave me the opportunity to work in her lab, guiding me from Computer Engineering to Bioinformatics and Computational Biology. Moreover, she founded QBM Graduate School, which I thank for all the support during these 4 years.

I would like to thank Prof. Dr. Roland Beckmann to have stepped in and helped during difficult times, and for the time spent on my thesis. Also, I would like to thank Dr. Stefan Canzar and Prof. Dr. Erwin Frey for their time as TAC committee members, examination committee members and for their guidance, and I would like to thank Prof. Dr. Julian Stingele, Prof. Dr. Karl Peter Hopfner and Prof. Dr. Klaus Förstemann for their time as examination committee members.

A special ‘thank you‘ goes to Andrea Storti and Marta Bozek. Without them and their help I would not have been able to write this thesis. Also, I would like to thank Ulrich Unnerstall for the help during these 4 years, Stefano and Zhan for our post-lunch breaks, and the entire Gaul group for the time spent together.

Another sincere ‘thank you‘ goes to Stefano Turolla, former Linux System Administrator of the Gene Center. Among other things, he helped us establishing our HPC cluster, that sensibly reduced computation time and, as a consequence, the time needed to complete the project.

I would like to start my non-job-related ‘thanks‘ by thanking my parents Rossano and Fernanda, who have always helped me and supported me in every possible way, and especially during these 4 years as a PhD student.

The second ‘thanks‘ goes to the friends in my hometown, and in particular to Francesco, Marco, Luigi, Riccardo, Irene, Elisabetta, Beatrice, Lavinia, Federico, Marilena, Chiara, Simona, Giuseppe, Andrea and Clarissa. It is nice that everytime I visit home, we are always able to meet and have a nice time together.

There is another group of friends with whom I meet every time I visit home that I want to thank, a bunch of ex-students of Engineering that likes to speak about politics and nerdy stuff. We always meet in the same place, and this place is a small village on the mountains called Acone. What is so special about this place? Well, they prepare a delicious pasta called *penne all‘aconese*, and we love it, so another ‘thank you‘ goes to this dish, to the ones who invented it, to the ones who prepare it and to the ones who eat it: Niccolo‘, Enrico, Michele, Pierpaolo, Alessio and Luca (good luck for your new job in New York City).

I also would like to thank the friends from Munich (Rahmi, Deniz, Saygin, Hazal, Alessandro, Ksenia, Umut, Gizem, Gökcen, Basak, Paulina and Mike). We have good times together, and we will keep it this way.

A mention and a ‘thanks‘ for another wonderful friendship go to Daniele (also known as Ing. Dante) and Emma: again congratulations for a new chapter in your life, and see you soon at the palace of the Queen in Edinburgh!

Finally, Valentina. I would need to write another thesis, only to thank you properly. So, just thanks.

Abstract

Proper regulation of transcription is fundamental in all aspects of life, from development to homeostasis. Gene regulation and regulation of transcription have been studied for decades, allowing us to understand many of their processes. Nevertheless, a complete knowledge of them is far from complete.

Ecdysone stimulation is a great paradigm to study regulation of transcription, because ecdysone triggers a very complex response cascade, in which hundreds of genes are heavily regulated by several transcription factors (TFs). These regulatory events can be detected with a great temporal resolution using DNase-seq and Position Weight Matrices (PWMs), and the corresponding transcriptional output can be detected at the same resolution using RNA-seq. These two data can be integrated to deepen our understanding of gene regulation.

Using ecdysone stimulation as a paradigm allows us to gather insights on its effect on the system. In fact, despite the vast amount of research on the steroid hormone ecdysone and its effects during the development of Drosophila melanogaster, a complete understanding of its mechanisms is still missing. Moreover, not all TFs belonging to the response have been characterized, and knowledge about them and their roles are still lacking.

In this thesis we characterized the ecdysone response of S2 cells using accessibility data and expression data, with an unprecedented temporal resolution. By integrating the two data, we described the relationship between ecdysone-responsive regulatory regions and transcription, showing that expression and regulatory response are strongly correlated. We defined a set of TFs involved in the response, and we measured their motif enrichment in responsive regulatory regions. Moreover, statistical modeling of the two data gave further insights on the ecdysone response, suggesting additional TFs involved in the response and their functionalities. Additionally, statistical modeling is able to predict expression from regulatory activity, giving insights on the relationship between regulatory regions and their target genes, and on which features are important to model transcriptional regulation.

On top of that, we gathered accessibility data in different tissues during larva-to-pupa transition, with an unprecedented temporal and spatial resolution. We characterized the chromatin landscapes, which are representative of the cell fates, and with a tissue-specific motif enrichment we identified new TFs that could be involved in the ecdysone response in vivo.

Chapter 1

Introduction

This chapter reviews the current state of knowledge about regulation of transcription, chromatin structure and steroid hormone ecdysone, which lay the foundation to our work. Moreover, it reviews the current efforts to model the relationship between gene expression, TFs input and chromatin accessibility to build regulatory networks.

1.1 Regulation of transcription

The genome is composed of coding regions and non-coding regions. The coding regions will be translated to proteins, whereas the non-coding regions hold the information that let proteins be created without errors and when they are needed. A subset of the non-coding regions is defined by cis-regulatory elements, which are sequences that control the transcription of genes. Cis-regulatory elements are bound by proteins called TFs, which regulate gene expression. Cis-regulatory elements can be further classified into promoters and enhancers. Promoters are located very close to the transcription start site (TSS) of a gene, usually immediately upstream, and they can be up to 1000 base pairs bp long. Promoters are responsible for regulating the strength of transcription, or equivalently how much mRNA should be transcribed from their regulated genes. Enhancers can be located up to several hundreds of thousands bp away from the TSS. However, the majority is usually located in proximity of the TSS, up to tens of thousands bp away. Enhancers can be both upstream and downstream of the TSS and they are responsible for the spatial and temporal regulation of transcription. Promoters and enhancers contain motifs, which are very short DNA sequences that TFs recognize in order to bind to the DNA. Motifs can be up to 20 bp, but are usually shorter. Each TF has a preferred motif to bind to. TFs bind to promoters and enhancers, and bound enhancers interact with bound promoters to regulate the transcription of genes. A visual representation of regulation of transcription is depicted in figure 1.

The DNA in the nucleus is wrapped around histones, and the complex formed by DNA and histones is called nucleosome. The position of nucleosomes along the genome and their fragility play a role in regulation of transcription. The fragility of a nucleosome is

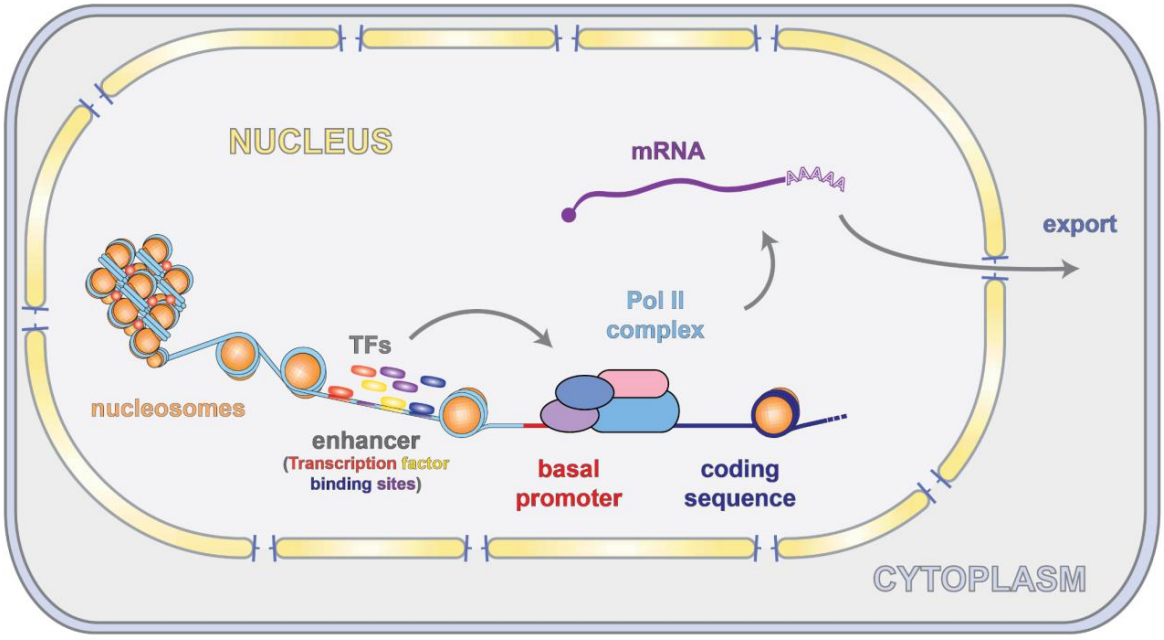

Figure 1: Schematic representation of transcription and its regulation. The Pol II complex binds to the promoter. TFs bind to enhancers, that interact with the Pol II complex at the promoter to regulate transcription. The transcription machinery transcribes the DNA coding sequence of genes into mRNA, which is exported from the nucleus into the cytoplasm, where it will be translated into a protein.

defined as the resistance to digestion from the enzyme MNase, and it is correlated with the tightness of the DNA around histones. In particular, if nucleosomes that are not fragile are positioned on promoters or enhancers, they typically impede the binding of other proteins, in particular TFs. This means that transcription can not happen, since the needed machinery can not be established. However, a class of TFs, called pioneer TFs, is able to bind to regulatory regions that are occupied by non-fragile nucleosomes, making these nucleosomes fragile and the regulatory regions available for binding to other TFs. In fact, if fragile nucleosomes are positioned on a promoter and the related enhancers, the transcriptional machinery can assemble and transcription can take place. Throughout this thesis, regions occupied by fragile nucleosomes will be denoted as *accessible*, whereas regions occupied by non-fragile nucleosomes will be denoted as *not accessible*. Several assays have been developed to measure the accessibility of the genome, with DNase-seq and ATAC-seq being the most widely used. Regions with high accessibility correspond to regulatory regions that are either active or ready to be active [Thomas et al., 2011].

1.1.1 Promoters

A promoter is a very small segment of DNA, 100-150 bp long, located immediately upstream of the TSS. The promoter can have different structures and functionalities, and it is composed of several elements that contribute to regulation in a combinatorial fashion. However, these structures are well defined and altering their DNA sequences changes the activity in a precise manner. For example a very important motif is the TATA box and its location, orientation and adjacent bases have an impact on its functioning. The complexity of the promoter suggests that one of its function is to determine the specificity and selectively communicate with enhancers [Smale and Kadonaga, 2003, Juven-Gershon et al., 2008].

The fragility of nucleosomes on the promoter allows Pol II and general TFs (GTFs) to be recruited to the DNA, in order to form the pre-initiation complex to start transcription. GTFs recognize regulatory elements in the promoter, independent from the regulated gene, whereas TFs recognize regulatory elements outside of the promoter. Proper TFs binding, proper chromatin structure and proper promoter organization are necessary to have high transcriptional activity, and even the alteration of one of them can destroy the correct functioning [Lubliner et al., 2015].

1.1.2 Enhancers

Enhancers are segments of DNA 100-1000 bp long that, when bound by TFs, interact with promoters to control spatio-temporal expression [Roeder, 1996]. Enhancers can be found several hundreds of thousands of bp away from the TSS, but thanks to the looping of the DNA they are brought in proximity to the promoters of the regulated genes. Enhancers can be found both upstream and downstream of the TSS, and particularly they can be found in intronic regions of genes [Levine, 2010].

The sequences of enhancers evolved during evolution, but their functionalities are conserved, probably because evolutionary changes are functionally compensated between each other. This is caused by the conservation of TFs motifs rather than the entire sequence of an enhancer [Arnold et al., 2013].

Given the importance of enhancers, it is fundamental to locate them, and several methods have been developed. One of them [Kvon et al., 2014] makes use of the Vienna Tiles, which is a set of transgenic flies where non-coding regions have been divided into 2kbp segments, and each segment equipped with a reporter indicating its activity. Reporter expression data have been gathered across 7 stages during embryo development, together with chromatin accessibility data. The expression of the reporters, which indicates activity of enhancers, correlates with high accessibility of the reported segments, suggesting that active enhancers can be detected by detecting regions with high accessibility.

1.1.3 TFs

TFs are proteins that are able to bind to DNA and to regulate transcription. They bind to enhancers and are responsible for the proper spatio-temporal expression of genes. Activators TFs have the effect of increasing transcription upon binding, whereas repressors TFs have the effect of decreasing it. The sequences of promoters and enhancers contain motifs for the TFs that have to bind there, and regulation of transcription happens with an interplay of activators and repressors binding to them [Stanojevic et al., 1991, Spitz and Furlong, 2012]. However, not every binding event is functional and chromatin structure is fundamental to direct binding of TFs [Li et al., 2008]. The current estimates of the number of TFs in Drosophila melanogaster range between less than 1000 and more than 2000 [Pfreundt et al., 2009, Shazman et al., 2013]. TFs are responsible for the proper development of an organism, for the response of an organism to the environment, for example response to heat shock, and for the response to hormones such as ecdysone in the case of Drosophila.

Each TF has a preferred sequence to bind called consensus sequence. However, TFs can also bind sequences that have mismatches from the consensus sequence, and each mismatch penalizes in a different way the binding. Generally, more mismatches mean that a TF binds more weakly to the TFBS. This phenomenon is called weak binding and it has been shown to play an important role in the regulation of gene expression [Tanay, 2006, Segal et al., 2008]. PWMs are the most employed tool to model binding of TFs to the DNA, and they will be described in 1.2.5.

1.1.4 Chromatin

Chromatin is the complex formed by DNA and histones, and one of its functions is to regulate gene expression. Histones are wrapped in DNA in two turns for a total of 147 bp, forming a complex called nucleosome. Nucleosomes are separated between each other from 20 to 60 bp of DNA, called linker DNA. Position and fragility are two important properties of nucleosomes, which contribute to transcriptional regulation. Fragility is defined as the resistance to digestion from MNase, and correlates with the tightness of the DNA around

histones. In particular, non-fragile nucleosomes generally hinder the binding of TFs. However, pioneer TFs are able to increase the fragility of nucleosomes. This phenomenon is called chromatin remodeling, and it allows the binding of other TFs to regulatory regions for proper spatio-temporal regulation of gene expression.

Chromatin structure plays a role during development. In fact, changes in chromatin accessibility have been observed during the development of Drosophila embryo, and accessibility identified regulatory regions that were experimentally validated. Moreover, clusters of accessible regions are located near genes that encode for TFs, and a correlation exists between changes in accessibility and mRNA expression patterns [Thomas et al., 2011]. Drosophila imaginal discs of different appendages share accessible DNA regulatory modules at a given stage along development, except for the loci that code for master regulators. In addition, open chromatin profiles change during development and such changes are coordinated between imaginal discs [McKay and Lieb, 2013].

Given the importance of chromatin structure for gene expression and development, characterization of chromatin states using several chromatin marks has been studied in human cells [Ernst and Kellis, 2010] and Drosophila melanogaster cells [Kharchenko et al., 2011].

1.2 Tools

1.2.1 ChIP-seq

ChIP-seq [Johnson et al., 2007] is a method developed to detect binding of TFs in vivo. First, proteins are cross-linked to the DNA, which is subsequently sheared, usually with sonication. Then, the protein of interest is immunoprecipitated using antibodies with attached beads. Subsequently, the precipitated segments of DNA are unlinked from the proteins, purified and sequenced. An enrichment of DNA segments mapped to a region in the genome indicates that the targeted TF was bound in such region.

ChIP-seq has been very useful to detect regulatory events, however it has some limitations. If the targeted protein binds unspecifically to the DNA, the DNA segments will be immunoprecipitated and sequenced. However, this binding does not have a regulatory effect, giving rise to a false positive. Another limitation is that antibodies do not have a perfect efficiency at immunoprecipitating proteins, and developing antibodies for some proteins may be difficult, giving rise to false negatives. Moreover, each target TF needs a different ChIP-seq experiment, making its usage unfeasible for projects that involve more than a moderately high number of proteins to study.

1.2.2 DNase-seq

The DNase-seq protocol starts with the digestion of nuclei with DNase I, an enzyme that preferentially cuts the DNA in accessible regions. Very long fragments that pollute the detection of accessible regions are discarded with a size selection step. Subsequently, adaptors

are ligated to the ends of the fragments and sequenced. An enrichment of DNA segments mapped to a region in the genome indicates that such region is accessible. DNase-seq steps are represented in figure 2.

Almost 40 years ago, it was observed that hypersensitivity to DNase I is a function of open chromatin, and it was suggested that hypersensitivity holds only in the appropriate cell type or developmental stage [Wu, 1980]. With the development of high throughput methods, chromatin accessibility could be characterized genome-wide, using DNase-chip and DNase-seq [Boyle et al., 2008]. A few years later, DNase I hypersensitive sites (DHSs) were mapped for the entire human genome [Thurman et al., 2012] and for the entire Drosophila genome [Thomas et al., 2011].

DNase-seq was used, together with another assay called FAIRE-seq, to measure accessibility on several cell types. DHSs identify regulatory elements that define cell type, and open regulatory elements form clusters close to each other that could be needed to maintain cell identity. Moreover, open chromatin that is cell type specific is close to expressed genes in such cell type, and DHSs identify the majority of bound TFBSs [Song et al., 2011]. A similar result was obtained in [Kaplan et al., 2011], where the authors presented a quantitative model of the mechanism that controls patterns of TFs binding in early Drosophila embryo development. By incorporating accessibility data in their model the performances greatly improved, meaning that in regions of open chromatin, binding can be predicted almost exclusively from the sequence specificity of TFs calculated using PWMs, making it possible to target every TF in a single assay and not having to rely on multiple ChIP-seq experiments.

1.2.3 ATAC-seq

ATAC-seq is an accessibility assay that was recently developed [Buenrostro et al., 2013], and it uses a different enzyme than DNase-seq. This enzyme is called Tn5 and it is a transposase, which is preloaded with the adaptors that have to be ligated for sequencing. Thanks to this, in the ATAC-seq protocol cleavage and insertion of adaptors happen at the same time. This is the main difference with the DNase-seq protocol (figure 2). Simultaneous cleavage and adaptors insertion has two advantages. First, it requires less starting material compared to DNase-seq. A successful ATAC-seq experiment can be performed using 500-50000 cells, whereas DNase-seq requires millions to tens of millions of cells. Second, the execution of the experiment is shorter. The preparation time is less than a day for ATAC-seq, whereas DNase-seq requires 3 days. For these reasons, ATAC-seq is well suited for in-vivo applications, where collection of cells may be difficult. Moreover, the analysis of fragment sizes from an ATAC-seq experiment showed that it can also be used to detect nucleosomes. This was also achieved for DNase-seq by applying an additional size selection step to fragments before sequencing [Vierstra et al., 2014].

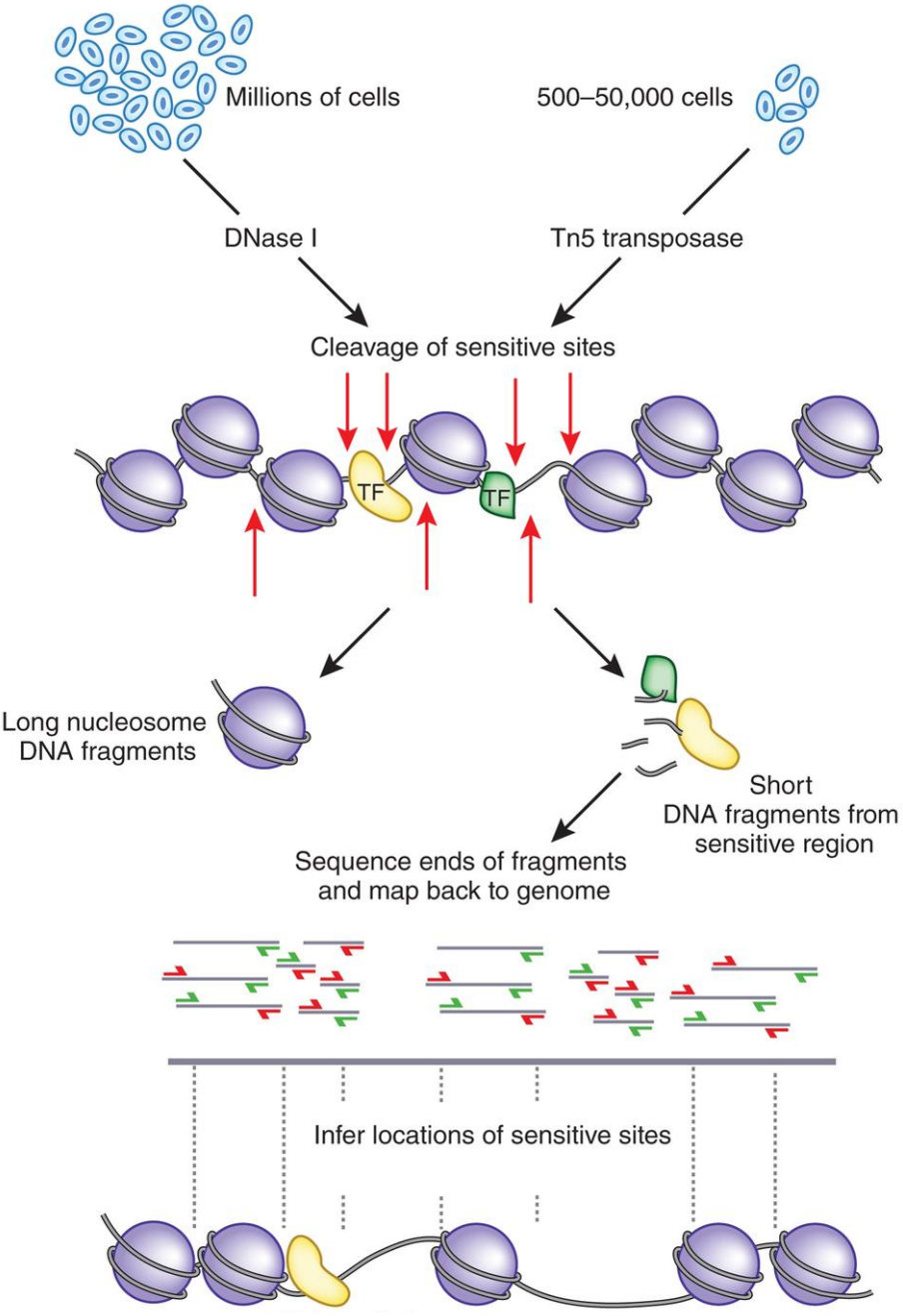

Figure 2: Schematic representation of DNase-seq and ATAC-seq protocols. The starting material is digested with DNase I (DNase-seq) or Tn5 transposase (ATAC-seq). Long nucleosomal fragments are discarded and short fragments from accessible regions are sequenced and mapped to a reference genome. As a last step, enrichment of mapped fragments is detected to find accessible regions. This task is referred to as *peak calling*. Figure adapted from [Raj and McVicker, 2014].

1.2.4 Digital genomic footprinting

Once open chromatin regions have been identified, it is desirable to identify bound TFBSs that lie within these regions. Digital genomic footprinting assumes that bound TFs protect the DNA from DNase I and Tn5 transposase cleavage, leaving a footprint in the distribution of cut sites. Searching for these footprints would allow the detection of bound TFBSs of all TFs without having to rely on PWM-based predictions of TFBSs, or without having to perform hundreds of ChIP-seq experiments. Several tools have been developed to detect footprints [Hesselberth et al., 2009, Pique-Regi et al., 2011, Cuellar-Partida et al., 2011, Neph et al., 2012, Piper et al., 2013, Sung et al., 2014, Gusmao et al., 2014, Piper et al., 2015], mostly on human cell lines.

However, in Drosophila footprint predictions poorly overlap with validated TFBSs (observation by Marta Bozek and Andrea Ennio Storti). In addition, there are three known issues with footprints detection. First, the residence time of a protein on the DNA correlates with the depth of the observed footprint [Sung et al., 2014]. This means that only TFs with a long residence time can be detected. Second, the enzymatic activity observed at TFBSs is dependent on the enzyme used to perform the chromatin accessibility experiment [Sung et al., 2014]. This means that the distribution of cut sites at TFBSs obtained with DNase-seq is different from the distribution of cut sites at TFBSs obtained with ATAC-seq. This is in contrast with previous assumptions that the distributions of cut sites reflected properties of the bound TF such as the contact mode [Neph et al., 2012]. Third, the distribution of cut sites at TFBSs is identical between digested chromatin and digested deproteinized genomic DNA for some TFs, whereas for other TFs the distributions are different [He et al., 2014]. This means that the enzymes are not cutting the DNA randomly, but that they have sequence biases [Koohy et al., 2013, Madrigal, 2015].

1.2.5 PWMs

Several models have been developed to represent DNA binding sites and TFs preferences [Stormo, 2000], but PWMs [Stormo et al., 1982] are by far the most widely used. In the case of DNA, a PWM has 4 rows and the number of columns is the number of bp of the TF binding motif. Each entry in a PWM can represent the frequency, the probability or the log-likelihood of observing a particular nucleotide at a particular position. Pseudocounts can be thought of as artificial observation added to the data, and are usually employed for at least two reasons: first, to avoid having zeroes or not finite values in the matrices, second, to avoid biases and to avoid having overspecific matrices due to a small sample size [Nishida et al., 2008]. Once a PWM to model the binding of a TF to the DNA has been defined, it is possible to use it to find putative TFBSs) by simply applying a threshold on the score obtained for the sequence at hand [Stormo, 2000]. The main limitations of PWMs are that the contribution of each bp to the final score is additive [Stormo, 2000], and that the calculation of each column is independent from the others. However, these limitations do not severely affect the calculation of TF affinities, therefore PWMs remain a good approximated representation of TFs preferences.

As it is fundamental to estimate binding preferences of TFs using PWMs, several methods have been developed, among which protein binding microarrays [Badis et al., 2009], bacterial one hybrid [Noyes et al., 2008], high throughput SELEX [Jolma et al., 2013] and HiP-FA [Jung et al., 2018]. The information has been gathered in several databases, among which Fly Factor Survey [Zhu et al., 2010] and JASPAR [Sandelin et al., 2004].

By using PWMs as main components, several methods and algorithms have been proposed, for example to count PWM occurrences in a sequence and to discriminate set of sequences with high counts and low counts of a PWM [Sinha, 2006], to predict the number of bound TFs to a sequence, or in other words to predict TF affinities to a sequence [Roider et al., 2006] and to predict expression using regulatory sequences [Segal et al., 2008].

1.3 The steroid hormone ecdysone

Ecdysone is a steroid hormone responsible for several morphological and behavioral changes in Drosophila, in particular it is responsible for the metamorphosis. The concentration of ecdysone varies during development and several pulses of concentration can be measured during Drosophila development. Two main pulses govern the metamorphosis with very precise timings, in particular the larval to prepupa and the prepupa to pupa transition, triggering a complex response cascade (figure 3).

To be functional, ecdysone binds to its receptor EcR-USP, an heterodimer formed by the TFs ecdysone receptor (EcR) and ultraspiracle (USP). The formed complex then binds to its TFBSs in the genome and triggers a complex transcriptional response. Nevertheless, according to the Ashburner model, the responding genes can be divided in two major groups: early genes, those that respond directly to the EcR-USP complex, and the late genes, those that respond to the early genes [Hill et al., 2013]. The majority of early genes code for TFs, and some of them were characterized: br [Chao and Guild, 1986], Eip74EF [Burtis et al., 1990], Eip75B [Segraves and Hogness, 1990] and EcR [Koelle et al., 1991]. Other known early genes are h, vri and Hr4, which are necessary for differentiation in the ecdysone response [Gauhar et al., 2009], and Eip78C, Hr39 and Hr3 [Huet et al., 1995]. Early genes repress themselves and induce late genes, while EcR-USP represses late genes [Baehrecke, 1996].

The timing of the isoforms of the main early TFs can be divided in two classes, namely class I with immediate response and class II with response at the peak of ecdysone, with br belonging to both classes [Karim and Thummel, 1992]. At the onset of metamorphosis, br has a key role in determining the stage specificity and the genetic hierarchy of the ecdysone cascade [Karim et al., 1993] and directly mediates the temporal and tissue specific response [Kalm et al., 1994]. Moreover, br has an isoform switch from br-Z2 to br-Z1, br-Z3 and br-Z4, with the latter 3 probably having a functional redundancy [Mugat et al., 2000].

The response has very different effects on the tissues of larvae, with two extreme effects: programmed cell death in larval tissues and differentiation in imaginal discs. Indeed, this is reflected in differential expression of genes in tissues at the onset of metamorphosis [Li and White, 2003]. Tissues with a different response to ecdysone express different EcR

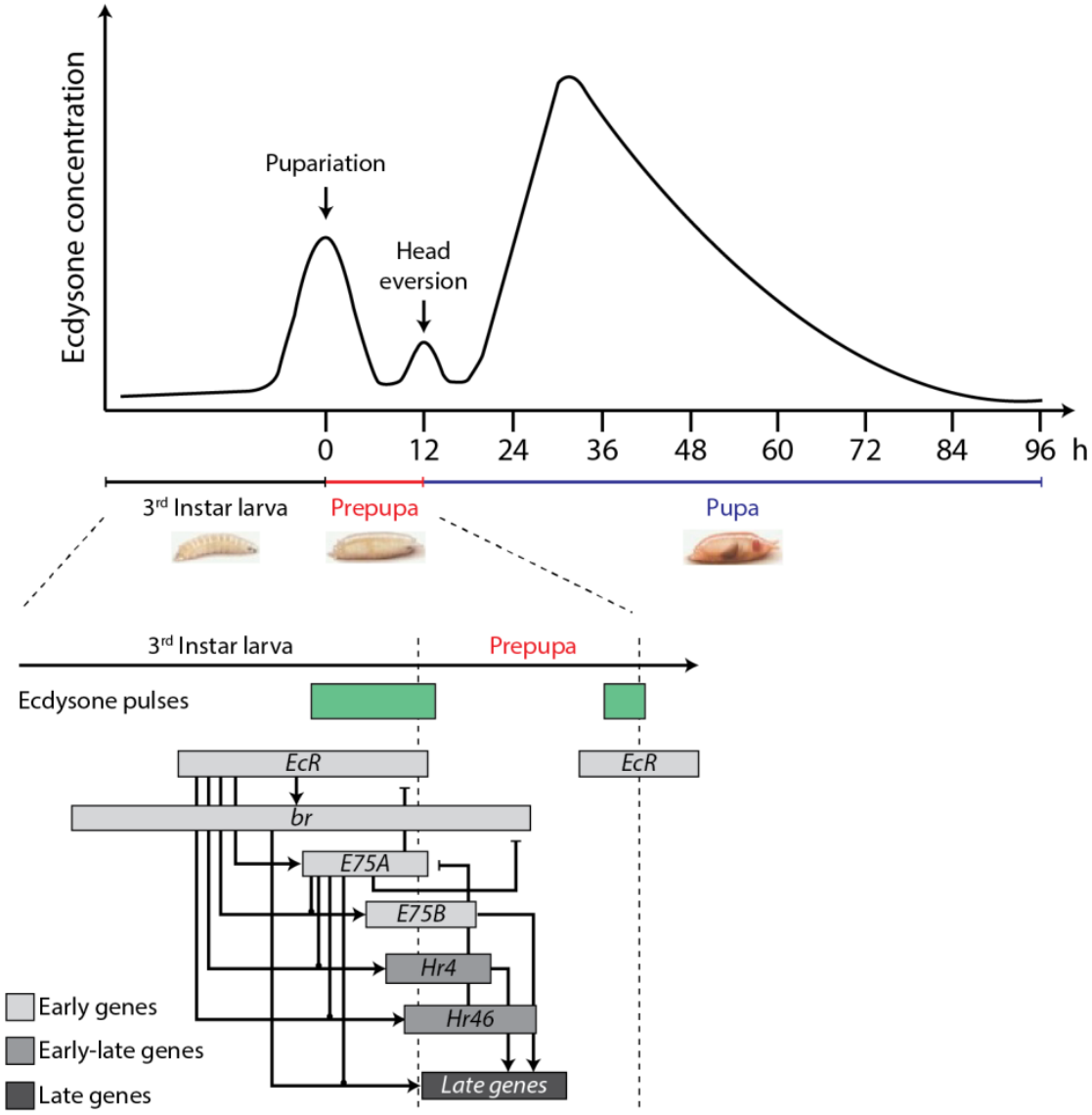

Figure 3: Schematic representation of ecdysone pulses during metamorphosis and its response cascade.
Upper part of the figure: measured ecdysone concentration from the larval stage to the pupa stage. Lower part of the figure: representation of the ecdysone response cascade. Lower part of the figure adapted from [Ou and King-Jones, 2013].

isoforms [Talbot et al., 1993], making the regulation of their expression an important switch during development [Robinow et al., 1993]. For example, in the central nervous system (CNS), some neurons will die during metamorphosis and some will survive and differentiate. Death of neurons is caused by hyperexpression of the isoform EcR-A, whereas expression of isoform EcR-B1 and subsequent switch to EcR-A, but not hyperexpressed, causes neurons to maturate [Truman et al., 1994]. Similar suggestions about the regulation of isoforms have been made for Eip74EF [Thummel et al., 1990] and for br. In particular, different tissues express different levels and combinations of br isoforms and this may characterize gene expression of each tissue, indicating that br has a global regulatory function along metamorphosis [Emery et al., 1994].

In S2 cells, activation of enhancers after ecdysone stimulation depends on a combination of motifs, and this combination is specific to S2 cells. In particular, EcR motifs are coupled with srp motifs, therefore srp may have an important role in the ecdysone response. Enhancers activated upon stimulus are not accessible before induction. By contrast, repression of enhancers after ecdysone stimulation seem independent from EcR motifs, and it may rely on motifs of other TFs. In particular, it appears to involve Eip74EF motifs. Strong induction of genes is mediated by multiple enhancers that are induced by ecdysone, whereas repression due to ecdysone is not direct, but rather mediated by other TFs [Shlyueva et al., 2014].

1.4 Gene expression modeling efforts

Mathematical and statistical models have been developed over the years to understand regulation of gene expression. Several of them model the relationship between gene expression levels and scores that relates TFs affinities to genes. Usually, the functionality of TFs is suggested by some coefficients that the models assign to them. If the coefficient is positive, the related TF is suggested to be an activator, whereas if the coefficient is negative, the related TF is suggested to be a repressor.

ChIP data were considered as a binary information on binding. Instead, ChIP data are quantitatively informative, and ChIP measurements reflect the strength of TF binding [Tanay, 2006]. In particular, weak binding is important for gene regulation, for example to have weak gene expression [Segal et al., 2008], therefore it is important to consider it while modeling gene expression.

One of the first works that use ChIP models gene expression using ChIP-seq data of 12 TFs. To do so, for each gene the authors define a window centered on its TSS, and each ChIP-seq peak that fall into the window is assigned to the relative gene. Subsequently, for each measured TF, they sum all the assigned ChIP strength, exponentially down-weighting them with their distance to the TSS. At this point, for each pair of genes and measured TFs, they have a TF-gene score. They regress these scores with gene expression levels to get coefficients that suggest the functionalities of the 12 TFs [Ouyang et al., 2009]. Although using only 12 TFs to model gene expression is a limitation, this work introduces ideas such as target gene assignment using a window centered on the TSS, the definition of

the TF-gene scores and the down-weighting with the distance that are used in subsequent works, including this thesis.

A similar regression model is presented in [McLeay et al., 2012], where instead of ChIP-seq data the authors predict binding using PWMs for the 12 TFs. They use FIMO [Grant et al., 2011], with histone modification and chromatin accessibility priors [Cuellar-Partida et al., 2011], to get binding scores to use in place of ChIP-seq measurements in the calculation of TF-gene scores. This work is worth a mention, because it shows that ChIP-seq experiments are not necessary to model gene expression, and finding active TFBSs using PWMs in open chromatin region suffices, in agreement with [Kaplan et al., 2011].

A different strategy to estimate TF-gene scores is presented in [Natarajan et al., 2012]. To assign TFs to genes, first the authors assign DHSs regions to the nearest TSS. Then, for each pair of genes and TFs, they compute the TF-gene score using a sliding window over the assigned DHSs, and by taking the maximum TFBS score. They use these scores in a logistic regression model to classify gene expression that is tissue specific, and the largest estimated coefficients are used to suggest TFs that act as regulators in each tissue. Even though this work showed that TF-gene scores can have alternative definitions, taking only the maximum TFBS score ignores weak binding. Moreover, target gene assignment based on the nearest TSS could be too restrictive when DHSs are equidistant to TSSs.

Another model is presented in [Blatti et al., 2015]. Using published PWMs, the authors score the Drosophila genome for TFs affinities using their algorithm Stubb [Sinha et al., 2003]. Then, they incorporate accessibility measurements, defining a *motif + accessibility* score, which is able to approximate very well experimental ChIP data, suggesting that it is possible to use accessibility and motifs instead of planning expensive multi-TFs ChIP experiments. Subsequently, they use this score, together with expression of TFs and scores of evolutionary conservation, to find important regulators of several expression domains in Drosophila embryo. They model the association between expression domains and enhancers using linear classification.

Building on [Blatti et al., 2015], the work presented in [Duren et al., 2017] models expression and accessibility to predict regulatory associations in contexts that are not present in the data. Assuming that a good genome annotation is available, together with coordinates of enhancers and associations with their target genes, and assuming protein-protein interaction data, the authors infer from observed expression and accessibility how each regulatory element interact with transcriptional regulators to regulate expression of target genes, from which they exclude TFs. The expression of a target gene is modeled using TFs bound to regulatory elements that are associated to the target gene and active in the context under analysis, and information on the expression of bound TFs. The activity of a regulatory element is modeled using recruited chromatin regulators, their expression and the accessibility of the regulatory element itself. The recruitment of a chromatin regulator to a regulatory element is modeled using binding affinities to the regulatory element of TFs known to interact with the chromatin regulator, the expression and the specificity in the context under examination of such TFs. A limitation of this model is that they assume the availability of a good genome annotation, in particular coordinates of enhancers and associations with their target genes. For Drosophila melanogaster a big

annotation effort is undertaken by the modENCODE consortium [Roy et al., 2010], with more than 700 datasets comprising DNase-seq, ChIP-seq and RNA-seq measured across several developmental stages and in several cell lines. The modENCODE consortium was able to annotate more than 20000 putative regulatory regions by an integrative analysis of their data [Nègre et al., 2011]. However, annotations of model organisms are far from complete and this task remains a significant challenge.

Another model that regresses expression levels and TF-gene scores is presented in [Schmidt et al., 2016]. Active regulatory regions are scored using the method presented in [Roider et al., 2006], which has the advantage of scoring each region as a whole, including in the calculation also weak TFBSs and not only the ones that are above some user-defined threshold. TFs-genes scores are computed analogously to [Ouyang et al., 2009], incorporating the accessibility of the regulatory element in the calculation. The authors use elastic net [Zou and Hastie, 2005] to regularize the estimation of the regression coefficients used to suggest the functionalities of TFs. The elastic net improves the interpretability of the estimated regression coefficients, because it sets to zero coefficients of TFs that are not predicted to be involved in gene regulation. Moreover, it keeps non-zero coefficients as small as possible, while distributing the weight between coefficients that are correlated, which correspond to TFs acting together to regulate transcription.

A follow-up of this work uses logistic regression to suggest TFs that are responsible for differential regulation between different conditions [Durek et al., 2016]. The features are defined as the ratio of TF-gene scores between different conditions, whereas the response variable is a binary variable that tells whether a gene is upregulated or downregulated. Again, the authors use elastic net regularization to estimate the coefficients. The ratio between TF-gene scores represents the change of affinity between conditions that a TF has with respect to a regulated gene.

1.5 Aim of the thesis

One of the research interests of the Gaul lab is understanding the mechanisms of gene regulation, in particular at the transcriptional level. Regulatory events are driving transcription, therefore it is necessary to pinpoint them as accurately as possible. This can be done using ChIP-seq, with an experiment for each TF under study, or with DNase-seq, with a single experiment to detect regulatory events genome-wide.

DNase-seq data can be used at two levels of resolution. At the coarser level it is used to detect open regulatory regions, and regulatory events are pinpointed using PWMs. At the finer level it is used to directly pinpoint regulatory events, using digital genomic footprinting. However, several issues of digital genomic footprinting make its usage not straightforward. The first aim of this thesis is to understand to which extent and at which resolution DNase-seq data can be used to understand regulation of transcription.

The ecdysone response is a valuable paradigm to study regulation of transcription. Ecdysone triggers a very complex cascade with hundreds of TFs involved and with thousands of regulatory events heavily regulating transcription. This is reflected in the extreme morphological and behavioral changes that happen in Drosophila after the ecdysone pulses. DNase-seq data gathered along the time course is used to localize regulatory events with an unprecedented spatio-temporal resolution, while RNA-seq data is gathered to measure the output of transcription along the time course. The second aim of this thesis is to integrate these two data to test how well they are correlated, and whether expression can be predicted from the regulatory events as mapped by chromatin accessibility.

Despite decades of research on the ecdysone response, our knowledge of all its mechanisms and effects is far from complete. The genes responding to ecdysone can be grouped in two classes: early genes, which respond directly to ecdysone, and late genes, which respond to the early genes. However, only a few TFs have been assigned to these classes, and knowledge of all the players involved in the response cascade is lacking. By using the ecdysone response as paradigm to study regulation of transcription, it is also possible to deepen the understanding of the ecdysone response itself. Therefore, by integrating and modeling accessibility and expression data, the third aim of this thesis is to characterize the ecdysone response and suggest new players involved in the cascade.

Chapter 2

Methods

2.1 Data acquisition

DNase-seq and nascent RNA-seq data were gathered at 6 different time points: untreated controls (UTC), 1 hours, 2 hours, 4 hours, 8 hours and 12 hours after stimulation. Nascent RNA-seq data were gathered by Katja Frühauf. Reads mapping and fragments per kilobase of transcript per million mapped read (FPKM) counting were performed by Thomas Walzthöni. DNase-seq data were gathered by Andrea Ennio Storti and Marta Bozek.

ATAC-seq data were gathered in 4 different tissues: eye discs (ED), wing disc (WD), salivary glands (SG) and CNS. Tissues were selected to be representative of the entire range of responses to the ecdysone pulse: ED and WD are associated with survival and differentiation, SG is associated with programmed cell death and CNS is associated with a mixed fate. Data for each tissue were gathered in 3 different stages: early 3rd instar larva (E3IL), late 3rd instar larva (L3IL) and white prepupa (WPP). Stages were selected to encompass the ecdysone pulse responsible for pupariation: E3IL just before the pulse, L3IL near the peak of the pulse and WPP after the pulse. ATAC-seq data were gathered by Andrea Ennio Storti.

2.2 Sequencing and mapping of reads

Libraries were sequenced on an Illumina GenomeAnalyzer IIx to have 50 bp paired-end reads. Sequencing was performed by LAFUGA at the Gene Center of the LMU Munich. Sequenced reads have been demultiplexed using the provided barcodes, the Illumina index read and the tool *Illumina Demultiplex*, available in the customized installation of Galaxy [Giardine et al., 2005] of the Gene Center. For each sample, adaptors were trimmed using the tool *Clip adaptor sequence* available in the customized installation of Galaxy of the Gene Center, with settings

- Seed 5
- Mismatches in adaptor 0

- Minimum length after clipping 0
- Output clipped and non-clipped one file

Trimming of adaptors was necessary because some fragments were shorter than the sequencing length, causing adaptors to be sequenced. As a consequence, without trimming of adaptors these read would not have mapped to the reference genome. The files for each sample were downloaded from the customized installation of Galaxy of the Gene Center, and mapped locally using Bowtie 2 [Langmead and Salzberg, 2012] version 2.2.9 with the following bash command:

```
bowtie2 --quiet --local --very-sensitive-local --threads 16 --mm -x /opt/
   ↪ bowtie2-2.2.9/indexes/dm3
```

The index was downloaded from `ftp://igenome:G3nom3s4u@ussd-ftp.illumina.com/Drosophila_melanogaster/UCSC/dm3/Drosophila_melanogaster_UCSC_dm3.tar.gz`. Mapped reads were filtered for mapping quality and proper pairing using Samtools [Li et al., 2009] version 1.3.1 with the following bash command:

```
samtools view -f 0x3 -q 10
```

The parameter *-f 0x3* was used to retain only reads that are paired and that are mapped in a proper pair. The parameter *-q 10* was used to retain only reads with a *MAPQ* score equal or bigger than 10. The *MAPQ* score is assigned from Bowtie 2 to each read, and it represents the confidence of having mapped the read in the right place in the reference genome. Reads that map in multiple places in the genome are assigned a *MAPQ* score of 0 or 1, therefore they were filtered. Filtered reads were sorted and indexed using Samtools version 1.3.1.

2.3 Peak calling

Peaks were called on each sample using MACS2 [Zhang et al., 2008b] version 2.1.1, using a gDNA sample as control, with the following bash command:

```
macs2 callpeak --keep-dup all -q 0.01 --nomodel --shift -100 --extsize 200
   ↪  -f BAM -g dm -B
```

MACS2 was chosen because, even though it was developed for ChIP-Seq data, it works well also with DNase/ATAC-Seq data. Its functioning is defined as follows. The parameters *–nomodel –shift -100 –extsize 200* were used to specify a smoothing window of 200 bp, and to center it on the 5‘ ends of mapped reads. After a smoothed pile up of 5‘ ends is computed, in a first pass enrichment of reads is scored using a Poisson distribution with a lambda parameter estimated from a genome-wide background. In a second pass, the score is refined using a local lambda parameter estimated from a local background defined using a naked DNA control sample.

2.4 Detection of differential peaks

2.4.1 S2 cells

Using DESeq2

Time points were compared against UTC in a pairwise fashion. A common set of peaks for a pair of time points was derived by taking only peaks that are present in both duplicates for the same time point, using BEDTools [Quinlan and Hall, 2010] version 2.26.0 and the sub-command *intersect*. Subsequently, the sets of remaining peaks in both time points were unified using BEDTools version 2.26.0 and the sub-command *merge* to obtain a common set of peaks for a pair of time points. The cut sites in each peak in the common set were counted for both time points using BEDTools version 2.26.0 and the sub-command *coverage*, with the following bash command:

```
bedtools coverage -sorted -counts
```

The differential peaks were called using the R/Bioconductor [R Core Team, 2018, Huber et al., 2015] package DESeq2 [Love et al., 2014] with an FDR threshold of 0.01.

Using ImpulseDE2

A common set of peaks for the entire time series was derived by taking only peaks that are present in both duplicates for the same time point, using BEDTools version 2.26.0 and the sub-command *intersect*. Subsequently, the sets of remaining peaks in each time point were unified using BEDTools version 2.26.0 and the sub-command *merge* to obtain a common set of peaks for the entire time series. The cut sites in each peak in the common set were counted for each time point using BEDTools version 2.26.0 and the sub-command *coverage*, with the following bash command:

```
bedtools coverage -sorted -counts
```

The differential peaks were called using the R/Bioconductor package ImpulseDE2 [Fischer et al., 2017] with an FDR threshold of 0.01. The differential peaks were classified into the classes *Transition Up*, *Transition Down*, *Transient Up*, *Transient Down* by ImpulseDE2.

2.4.2 Larvae

Differential peaks in larvae were called using MACS2 version 2.1.1 using the sub-command *bdgdiff* and following the instructions on the page `https://github.com/taoliu/MACS/wiki/Call-differential-binding-events`. MACS2 was used instead of some more established differential analysis tool due to the lack of replicates [Steinhauser et al., 2016].

2.5 Detection of differential genes

2.5.1 S2 cells

Using DESeq2

Time points were compared against UTC in a pairwise fashion. The differential genes were determined using the R/Bioconductor package DESeq2 with an FDR threshold of 0.01, using counts provided by Thomas Walzthöni.

Using ImpulseDE2

The differential genes for the entire time course were determined using the R/Bioconductor package ImpulseDE2 with an FDR threshold of 0.01, using counts provided by Thomas Walzthöni. The differential genes were classified into the classes *Transition Up*, *Transition Down*, *Transient Up*, *Transient Down* by ImpulseDE2.

2.6 Assignment of target genes to peaks

2.6.1 Nearest TSS strategy

The association between peaks and candidate target genes was done based on the distance of the peaks from the TSSs of genes. Each peak was assigned to the gene with the nearest TSS. This operation was performed using the R/Bioconductor package ChIPseeker [Yu et al., 2015]. Further assignments were done using only expressed genes in the time course, or only differential genes in the time course.

2.6.2 Regions of influence strategy

The association between peaks and candidate target genes was done defining regions of influence for each gene, inspired by [McLean et al., 2010]. Let exons be numbered according to their genomic position in a chromosome, from the leftmost position to the rightmost position. Let s_i be the coordinate of the first bp of the first exon i of gene g, and let e_j be the coordinate of the last bp of the last exon j of gene g. Note that $i \leq j$. The region of influence ROI_g of gene g is defined as follows:

$$ROI_g = \begin{cases} [e_{i-1}, s_{j+1}] & \text{if } s_i - e_{i-1} \leq D \text{ and } s_{j+1} - e_j \leq D \\ [\alpha e_{i-1} + (1-\alpha)s_i, s_{j+1}] & \text{if } s_i - e_{i-1} > D \text{ and } s_{j+1} - e_j \leq D \\ [e_{i-1}, \alpha e_j + (1-\alpha)s_{j+1}] & \text{if } s_i - e_{i-1} \leq D \text{ and } s_{j+1} - e_j > D \\ [\alpha e_{i-1} + (1-\alpha)s_i, \alpha e_j + (1-\alpha)s_{j+1}] & \text{if } s_i - e_{i-1} > D \text{ and } s_{j+1} - e_j > D \end{cases} \tag{2.1}$$

where $0 < \alpha < 1$ and D is defined as the maximum distance between exons i, j of gene g and exons $i-1, j+1$ of adjacent genes to have ROI_g reach adjacent genes. All peaks

that overlapped ROI_g were assigned to gene g. Further assignments were done using only expressed genes in the time course, or only differential genes in the time course.

2.6.3 Window centered on TSS strategy

The association between peaks and candidate target genes was done using a window centered on the TSS of each gene, as seen in [Ouyang et al., 2009, McLeay et al., 2012, Schmidt et al., 2016]. Let TSS_g be the coordinate of the TSS of gene g. The window W_g of gene g is defined as:

$$W_g = [TSS_g - w, TSS_g + w] \tag{2.2}$$

where w is defined as the width of window W_g. All peaks that overlapped W_g were assigned to gene g.

2.7 Clustering of dynamics

2.7.1 Differential peaks

A profile composed of the $log_2(FC)$ of accessibility along the time course was assigned to each differential peak. Profiles were clustered using hierarchical clustering with Ward's minimum variance method [Ward Jr, 1963]. The dissimilarities between profiles needed for clustering were computed using the cosine distance. Let $\mathbf{a} = [a_1, a_2, \cdots, a_n]^T$ and $\mathbf{b} = [b_1, b_2, \cdots, b_n]^T$ be profiles of 2 differential peaks. The cosine similarity $cos\,(\theta)$ between $\mathbf{a}$ and $\mathbf{b}$ is defined as:

$$cos\,(\theta) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|_2 \|\mathbf{b}\|_2} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}} \tag{2.3}$$

Cosine distance is defined as:

$$d\,(\mathbf{a}, \mathbf{b}) = 1 - cos\,(\theta) \tag{2.4}$$

The cosine distance was chosen because it is invariant to the amplitude of the data, allowing profiles to be compared based only on their shape. Note that cosine similarity is equivalent to the PCC if the data is mean centered ($\overline{\mathbf{a}} = 0, \overline{\mathbf{b}} = 0$):

$$PCC = \frac{\sum_{i=1}^{n} (a_i - \overline{\mathbf{a}}) \left(b_i - \overline{\mathbf{b}}\right)}{\sqrt{\sum_{i=1}^{n} (a_i - \overline{\mathbf{a}})^2}\sqrt{\sum_{i=1}^{n} \left(b_i - \overline{\mathbf{b}}\right)^2}} \tag{2.5}$$

2.7.2 Differential genes

The profiles of differential genes were clustered as described in 2.7.1.

2.8 GO analysis

Enrichment and depletion analysis of GO terms was performed using the R package BOAT, developed by Ivo Zeller and available at
`https://github.com/zellerivo/BOAT---biological-ontology-analysis-tool`.
The hypergeometric distribution and a two-sided statistical test are used to calculate a p-value for enrichment or depletion of terms. A doubling p-value approach was chosen as two-sided statistical test for BOAT, instead of the commonly used minimum likelihood definition. Plots of the results of the analysis were done using a R script written by Ivo Zeller included in BOAT. Simultaneous visualization of enrichments across different conditions is enabled by such R script, together with the representation of the cardinality of the experimental set and the reference set. These features were developed in BOAT for better visualization of the outcomes of a GO analysis. The package BOAT and the plotting script were developed as part of the Master intership of Ivo Zeller in the Gaul lab. As a reference and for further details I refer to his internship report.

2.9 Measurement of similarity between sets of genes

The Jaccard index [Jaccard, 1901] was used to measure the similarity between sets of genes. Let A and B be sets. The Jaccard index is defined as

$$J(A,B) = \begin{cases} 1 \text{ if } A = \emptyset \text{ and } B = \emptyset \\ \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \text{ otherwise} \end{cases} \tag{2.6}$$

The Jaccard index assumes values between 0 and 1, with 0 representing no overlap between sets and 1 representing identical sets.

2.10 Distances between samples of larval tissues

To measure the differences of the chromatin structure between larval tissues, DESeq2 [Love et al., 2014] was used. First, a common set of peaks for all tissues in all developmental stages was obtained in the same way as described in 2.4.1. The sub-command *intersect* of BEDTools was not used because the data was gathered without duplicates. The cut sites in each peak in the common set were counted as described in 2.4.1. After running DESeq2, its functions *rlog*, *dist* and *plotPCA* were used to measure and plot the distances between samples.

2.11 Determination of relevant TFs

2.11.1 Used for motif enrichment in S2 cells

The set of Drosophila TFs was determined as follows. First, all the TFs listed in the database FlyTF [Pfreundt et al., 2009] were retrieved. Second, all the TFs that were annotated with the gene ontology (GO) [Ashburner et al., 2000, Consortium, 2016] term *GO:0003700*, that stands for *sequence-specific DNA binding transcription factor activity*, were retrieved using the R/Bioconductor package biomaRt [Durinck et al., 2009]. Subsequently, these two sets were merged. TFs that showed a differential peak on their promoter and differential expression were considered to have differential behaviour. Availability of PWMs of TFs that showed differential behaviour was checked in the PWMs databases FlyFactorSurvey [Zhu et al., 2010] and JASPAR [Sandelin et al., 2004]. Whenever available, PWMs measured in the Gaul lab with the HiP-FA method [Jung et al., 2018] were used.

2.11.2 Used for motif enrichment in larval tissues

The set of TFs was determined as described in 2.11.1, replacing the definition of differential behaviour since expression data were not available. TFs that showed a differential peak along stages on their promoter were considered to have differential behaviour.

2.11.3 Used as features in the models

The R/Bioconductor package MotifDb [Shannon and Richards, 2017] was used to gather all PWMs from published databases. In case a TF had more than one PWM, we followed the procedure described in the TEPIC [Schmidt et al., 2016] documentation available at `https://github.com/SchulzLab/TEPIC`, and chose the one with the smallest IC, which is defined as:

$$IC = -\frac{\sum_{i,j} P(i,j)\, log_2 P(i,j)}{M} \tag{2.7}$$

where M is the length of the motif, $i \in \{A, C, G, T\}$, $j \in \{1, ..., M\}$ and $P(i,j)$ is the probability of observing nucleotide i at position j. Then, we filtered the set of PWMs to retain only TFs that are expressed along the time course. A TF is considered expressed in the time course if it has a FPKM count greater than 1 in at least 1 time point. Whenever available, PWMs measured in the Gaul lab with the HiP-FA method were used.

2.12 Motif enrichment

The FASTA files containing the nucleotide sequences of the differential peaks where obtained using BEDTools version 2.26.0 with the sub-command *getfasta*, using the reference genome `ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r5.`

`53_FB2013_05/fasta/dmel-all-chromosome-r5.53.fasta.gz`. The FASTA files containing the control sequences were obtained shuffling the nucleotide sequences of the differential peaks, maintaining the dinucleotide frequencies, using the tool *fasta-shuffle-letters* from the MEME Suite [Bailey et al., 2009], version 4.11.2, with the following bash command:

```
fasta-shuffle-letters -kmer 2 -dna
```

Motif enrichment was performed using AME [McLeay and Bailey, 2010] from the MEME Suite, version 4.11.2, with the following bash command:

```
ame --method ranksum --scoring avg
```

The set of TFs that showed differential behaviour was considered. Informations on TFs were retrieved using FlyBase [Gramates et al., 2016].

2.13 Definition of TF-gene scores

Let $a_{p,j}$ be the affinity of TF j with peak p computed using TRAP [Roider et al., 2006], P_i be the set of peaks assigned to gene i, r_j be the expression level of TF j, s_p be the mean accessibility of peak p, $d_{p,i}$ be the distance between the center of peak p and the TSS of gene i and d_0 be a constant fixed at 5000 bp [Ouyang et al., 2009]. The TF-gene score $x_{i,j}$ for gene i and TF j, $i \in \{1, ..., G\}$, $j \in \{1, ..., T\}$ is defined as:

$$x_{i,j} = r_j \sum_{p \in P_i} a_{p,j} s_p e^{-\frac{d_{p,i}}{d_0}} \tag{2.8}$$

If replicates are available, the final TF-gene score $x_{g,i}$ is defined as the mean of the scores in each replicate:

$$x_{i,j} = \frac{1}{N} \sum_{k=1}^{N} x_{i,j}^k \tag{2.9}$$

where $k \in \{1, ..., N\}$ is the sample number.

2.14 Regularized linear regression

Let $x_{i,j}$ be the TF-gene score for gene i and TF j as defined in 2.13, y_i be the expression value of gene i and β_j be the regression coefficient of TF j. The following matrices are defined:

$$X = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & \cdots & x_{1,T} \\ 1 & x_{2,1} & \ddots & & & \vdots \\ \vdots & \vdots & & x_{i,j} & & \vdots \\ \vdots & \vdots & & & \ddots & \vdots \\ 1 & x_{G,1} & \cdots & \cdots & \cdots & x_{G,T} \end{bmatrix}, \qquad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_G \end{bmatrix}, \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_T \end{bmatrix} \tag{2.10}$$

The matrices X and y are log-transformed with a pseudocount of 1 and standardized by column to have zero mean and unitary standard deviation. Linear regression states that

$$y = X\beta + \epsilon \tag{2.11}$$

where ϵ is the normally distributed residual given X. The estimated coefficients are defined as:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{2G} \|y - X\beta\|_2^2 + \lambda \left[\alpha \|\beta\|_1 + \frac{(1-\alpha)}{2} \|\beta\|_2^2 \right] \tag{2.12}$$

where the first term of the sum is the negative log-likelihood, which is equivalent to the residual sum of squares, and the second term of the sum is the elastic net regularization [Zou and Hastie, 2005]. The R/Bioconductor package glmnet [Friedman et al., 2010] is used to estimate the regression coefficient $\hat{\beta}$.

2.15 Regularized logistic regression

Let $q_{i,j}$ be the ratios of the TF-gene scores defined in 2.13 between time point t and time point u:

$$q_{i,j} = \frac{x_{i,j}^t}{x_{i,j}^u} \tag{2.13}$$

$q_{i,j}$ are interpreted as the variation between time point t and time point u of the TF-gene score for TF j and gene i. Let y_i be 1 if gene i is upregulated between time point t and time point u, 0 if gene i is downregulated. Let β_j be the regression coefficient of TF j. The following matrices are defined:

$$Q = \begin{bmatrix} 1 & q_{1,1} & q_{1,2} & \cdots & \cdots & q_{1,T} \\ 1 & q_{2,1} & \ddots & & & \vdots \\ \vdots & \vdots & & q_{i,j} & & \vdots \\ \vdots & \vdots & & & \ddots & \vdots \\ 1 & q_{G,1} & \cdots & \cdots & \cdots & q_{G,T} \end{bmatrix}, \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_T \end{bmatrix} \tag{2.14}$$

The matrix Q is log-transformed with a pseudocount of 1 and standardized by column to have zero mean and unitary standard deviation. Let q_i be the i-th row of Q. Logistic regression states that

$$log\left(\frac{Pr\left(Y_i = 1|q_i\right)}{Pr\left(Y_i = 0|q_i\right)}\right) = \beta^T q_i \tag{2.15}$$

where $Y_i|q_i$ is Bernoulli-distributed with unknown probability p_i. The estimated coefficients are defined as:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} -\frac{1}{G} \sum_{i=1}^{G} \left[y_i \beta^T q_i - log\left(1 + e^{\beta^T q_i}\right) \right] + \lambda \left[\alpha \|\beta\|_1 + \frac{(1-\alpha)}{2} \|\beta\|_2^2 \right] \tag{2.16}$$

where the first term of the sum is the negative log-likelihood, and the second term of the sum is the elastic net regularization. The R/Bioconductor package glmnet is used to estimate the regression coefficient $\hat{\beta}$.

2.16 Cross validation

Cross validation is employed to prevent overfitting. An inner cross validation is used to fit α, λ and the regression coefficients $\hat{\beta}$, whereas an outer cross validation is used to assess performances.

First, a random 20% of the dataset is reserved as test data. On the remaining 80%, 10-fold cross validation is used to fit α, λ and the regression coefficients $\hat{\beta}$. Subsequently, the performances of the best fit are measured on the test data. The outer cross validation splitting is repeated 10 times, and the performance measurements are averaged across the 10 different splittings. The parameter λ is automatically estimated by the R/bioconductor package glmnet, whereas the parameter α is estimated using a grid search with step 0.01.

Finally, a model using 10-fold cross validation on the entire dataset is fitted for the interpretation of the estimated regression coefficients $\hat{\beta}$.

Chapter 3

Results

3.1 Characterization of the ecdysone response in S2 cells

3.1.1 S2 cells respond to ecdysone stimulation

S2 cells were derived in the early seventies from late stages of the Drosophila melanogaster embryo [Schneider, 1972]. In normal conditions, they proliferate and undergo cell division every 24 hours, whereas after ecdysone treatment they exit the cell cycle, stopping the proliferation. This is caused by cells starting to differentiate. Ecdysone stimulation is responsible for changes in morphology of S2 cells. In particular, they grow in size, they lose their round shape and they grow structures similar to filopodia [Frühauf, 2015]. Since S2 cells respond to the stimulus and they are easier to treat and harvest for the experiments than an in-vivo system, we decided to use them to study the ecdysone cascade.

To measure accessibility and expression changes upon ecdysone stimulation, we used DNase-seq and nascent RNA-seq data gathered at an unprecedented temporal resolution, as described in 2.1. To assess the quality of our DNase-seq data, we checked whether we could see an enrichment of cut sites in regulatory regions that have already been detected as accessible in our system. We took regulatory regions identified in S2 cells from [Arnold et al., 2013] using DNase-seq and STARR-seq, and we computed the average cut frequency using our DNase-seq data. As expected, we found that our data show enrichment of cut sites in those regions (figure A.1), so we are confident to identify functional regulatory regions.

A qualitative look at the data using a genome browser shows that expression changes are related to accessibility changes. Figure 4 shows expression and accessibility data measured at the br locus along time. It is possible to see that changes in accessibility, which comprise promoters of different isoforms and enhancers, correlates to changes in expression data over time. Since we have seen that our data is able to capture the response of the system to ecdysone stimulation, we measured the observed changes.

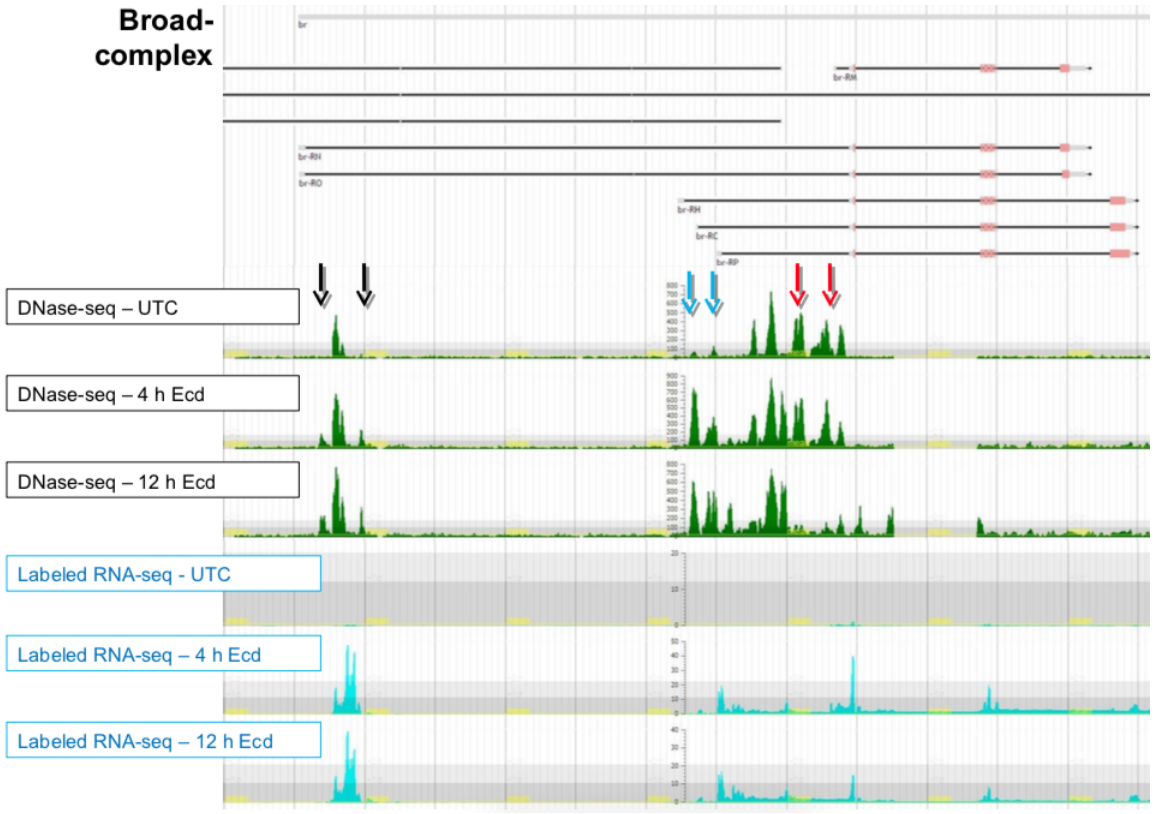

Figure 4: Example of DNase-seq and nascent RNA-seq data at the br locus. The arrows highlight regulatory regions that are opening or closing along the time course. It is possible to see that changes in accessibility are related to changes in expression.

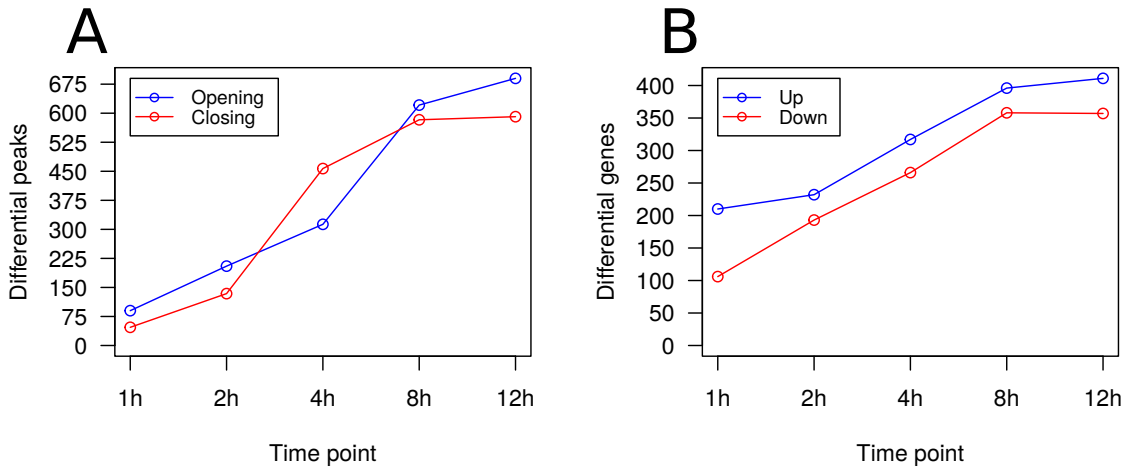

Figure 5: Number of differential peaks and differential genes along the time course.
(A) Number of differential peaks along the time course, separated by direction. 1h opening 90, closing 47. 2h opening 205, closing 134. 4h opening 313, closing 457. 8h opening 621, closing 583. 12h opening 690, closing 591. (B) Number of differential genes along the time course, separated by direction. 1h upregulated 210, downregulated 106. 2h upregulated 232, downregulated 193. 4h upregulated 317, downregulated 266. 8h upregulated 396, downregulated 358. 12h upregulated 411, downregulated 357.

3.1.2 Accessibility response and expression response are similar

To quantitatively explore the time course, we measured $log_2(FC)$ of differential genes and differential peaks as described respectively in 2.5.1 and 2.4.1. We separated differential genes and differential peaks by direction and by time point, and we counted them. The number of differential genes and differential peaks is steadily increasing along the time course, for both directions (figure 5). In the same figure and in figure 6 it is possible to see that the effect of ecdysone stimulation is not balanced along time. In particular, in the early time points and especially 1 hour after stimulation, the response is heavily imbalanced towards upregulation. In later time points the response becomes balanced. In the same figures it is also possible to see that the responses of accessibility and expression to the stimulation are similar. Moreover, proximal time points show a higher correlation, whereas distal time points are more scattered (figure 6).

Taken together, these results suggest that a correlation exists between chromatin modifications and gene regulation. Moreover, after stimulation the system has a strong immediate response followed by more stable adjustments. Motivated by these observed similarities, we investigated the relationships between differential peaks and gene regulation.

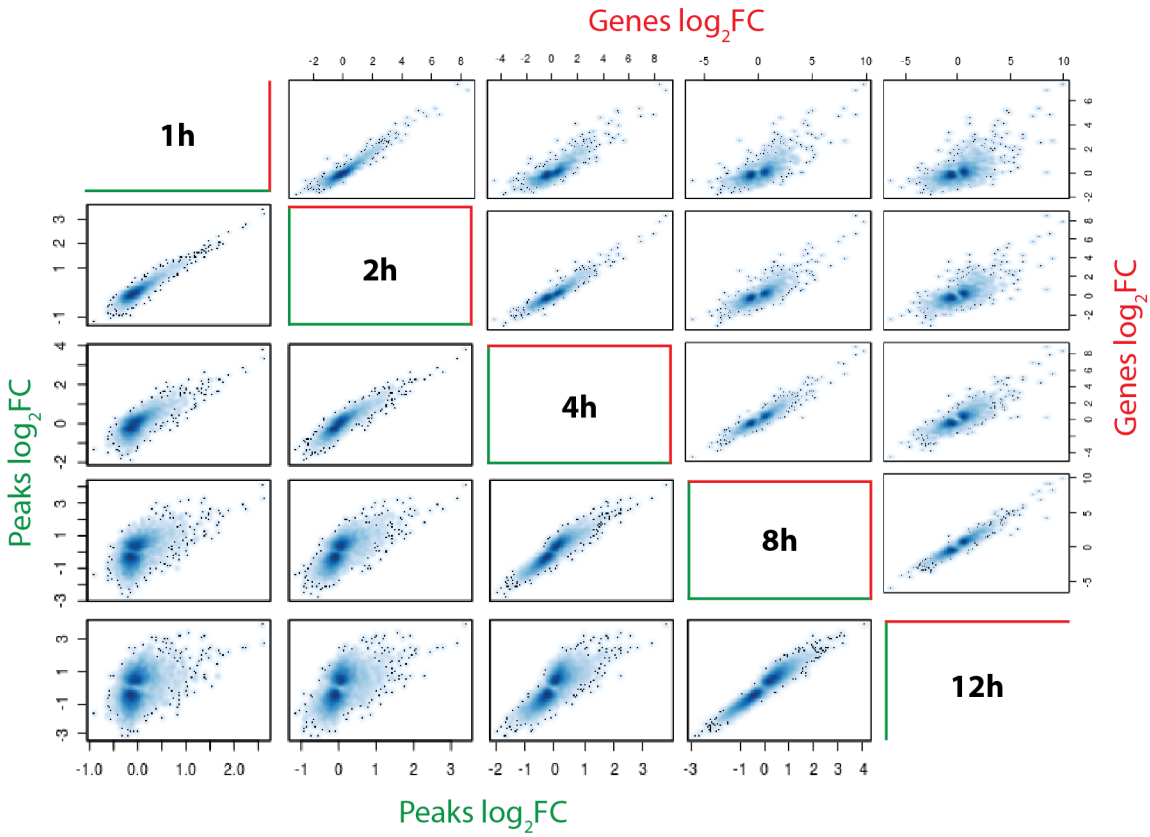

Figure 6: Scatterplots of $log_2(FC)$ of differential peaks and differential genes along the time course.
The $log_2(FC)$ of differential peaks (green) and differential genes (red) of a time point were compared against all the other time points. At early time points the distribution is skewed towards genes upregulation and chromatin opening. The response of accessibility and expression are similar along the time course.

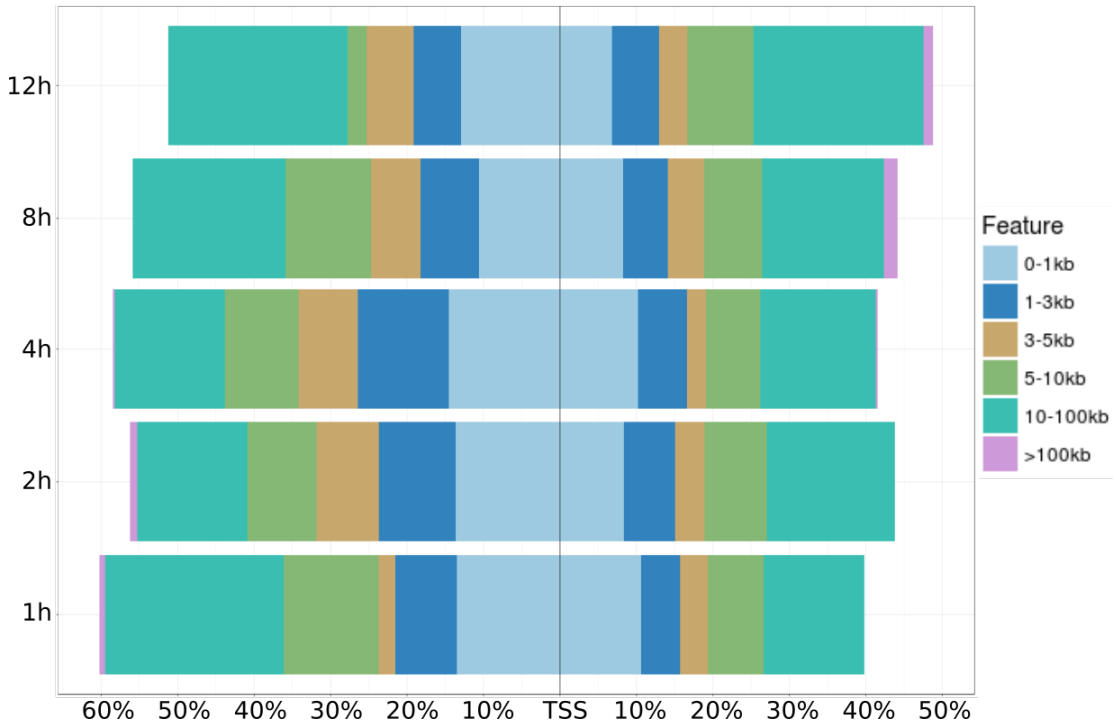

Figure 7: Distribution of distances between differential peaks and their nearest TSS for each time point.
Less than 2% of differential peaks are located more than 100k bp away from any TSS. At least 53% differential peaks are located within 10k bp from any TSS.

3.1.3 Assignment of target genes to peaks with the nearest TSS strategy is a good approximation

To do so, we assigned differential peaks to their candidate target genes. The correct association between enhancers and their target genes is not a trivial problem and it is still unsolved [Yao et al., 2015]. However, in Drosophila it appears that 88% of enhancers are either intragenic, directly upstream or directly downstream their target gene. This means that only 12% of enhancers have at least another gene between them and their target gene [Kvon et al., 2014]. Moreover, enhancers that are responsive to ecdysone regulate nearby genes [Shlyueva et al., 2014]. We checked the distribution of distances between differential peaks and their nearest TSS for each time point. Figure 7 shows that less than 2% of differential peaks are located more than 100k bp away from any TSS, and that at least 53% differential peaks are located within 10k bp from any TSS. If we exclude the 12h time point, this percentage raises to 66%.

Taken together, these data suggest that assigning candidate target genes to differential peaks using the distance between them is a good approximation of the real associations. For this reason, we decided to use the simple method of assigning as a target gene of a differential peak the gene that has the nearest TSS, as described in 2.6.1.

3.1.4 Direction of regulation correlates with direction of chromatin openness

To explore the relationship between differential peaks and their target genes, we correlated the $log_2(FC)$ between them. Figure 8 shows that at the beginning of the time course opening peaks are associated to upregulating genes and closing peaks are associated with downregulating genes, with very few exceptions. In later time points, more differential peaks are present and mixed associations appear, but the majority retains concordant directions. If we pool all the time points, the correlation is high (PCC 0.63) and 81.6% $log_2(FC)$ are concordant in sign. Closing peaks that have upregulated target genes are 14%, whereas only 4.1% opening peaks have downregulated target genes. Considering only enhancers in the analysis does not alter the results substantially (figure A.2). If we consider only promoters, correlation is generally higher (figure A.3). All the time points pooled have a PCC of 0.75 and 85.2% $log_2(FC)$ are concordant in sign.

Overall, this suggests a mechanism of regulation where the opening of regulatory regions generally has an activating effect on their target genes, whereas closing generally has a repressing effect. In promoters there are fewer exception to this mechanism than in enhancers. In particular, the exceptions are imbalanced towards closing enhancers that have upregulated target genes, suggesting a more sophisticated relationship between enhancers and their target genes.

3.1.5 Number of opening enhancers plays a role in gene upregulation

To analyze the relationship from the point of view of regulated genes, we checked whether the number of differential peaks associated with a gene plays a role in its regulation. We grouped genes having 1, more than 1 or more than 2 associated opening peaks. We did the same grouping with associated closing peaks. There is a significant shift towards higher values in the distribution of $log_2(FC)$ of genes with more than 1 opening peak, compared to genes with exactly 1 opening peak (figure 9). The shift is even more significant in the middle time points if we remove promoters from the analysis. By contrast, if we consider closing peaks, the distributions of $log_2(FC)$ of genes belonging to different groups are not significantly different, even after removing promoters.

This suggests that active regulatory regions have a joint effect in the upregulation of genes, whereas regulatory regions that become inactive do not collaborate in the downregulation of genes. Again, closing peaks have a more sophisticated role than simply shutting down gene expression.

3.1.6 Ecdysone stimulation triggers transient and permanent responses

To get further insights on the behavior of the system, we analyzed the dynamics along time of differential peaks and regulated genes. We clustered the profiles of the regulated genes

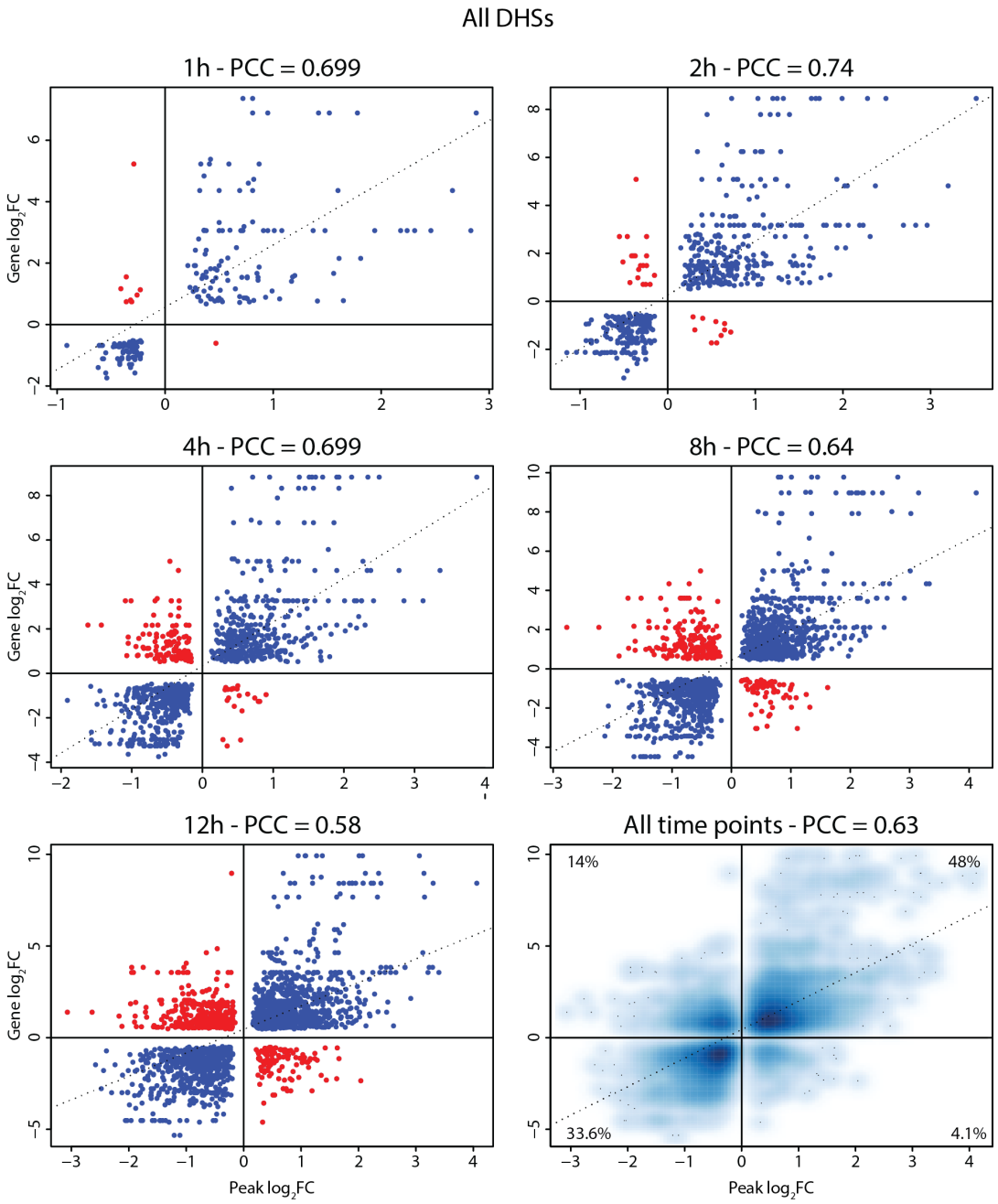

Figure 8: Scatterplots between $log_2(FC)$ of differential peaks and $log_2(FC)$ of their target genes.
For each time point, $log_2(FC)$ of differential peaks (x-axis) and $log_2(FC)$ of their target genes (y-axis) was correlated. Correlation values are shown above each plot. Blue dots: $log_2(FC)$ that agree in sign. Red dots: $log_2(FC)$ that do not agree in sign.

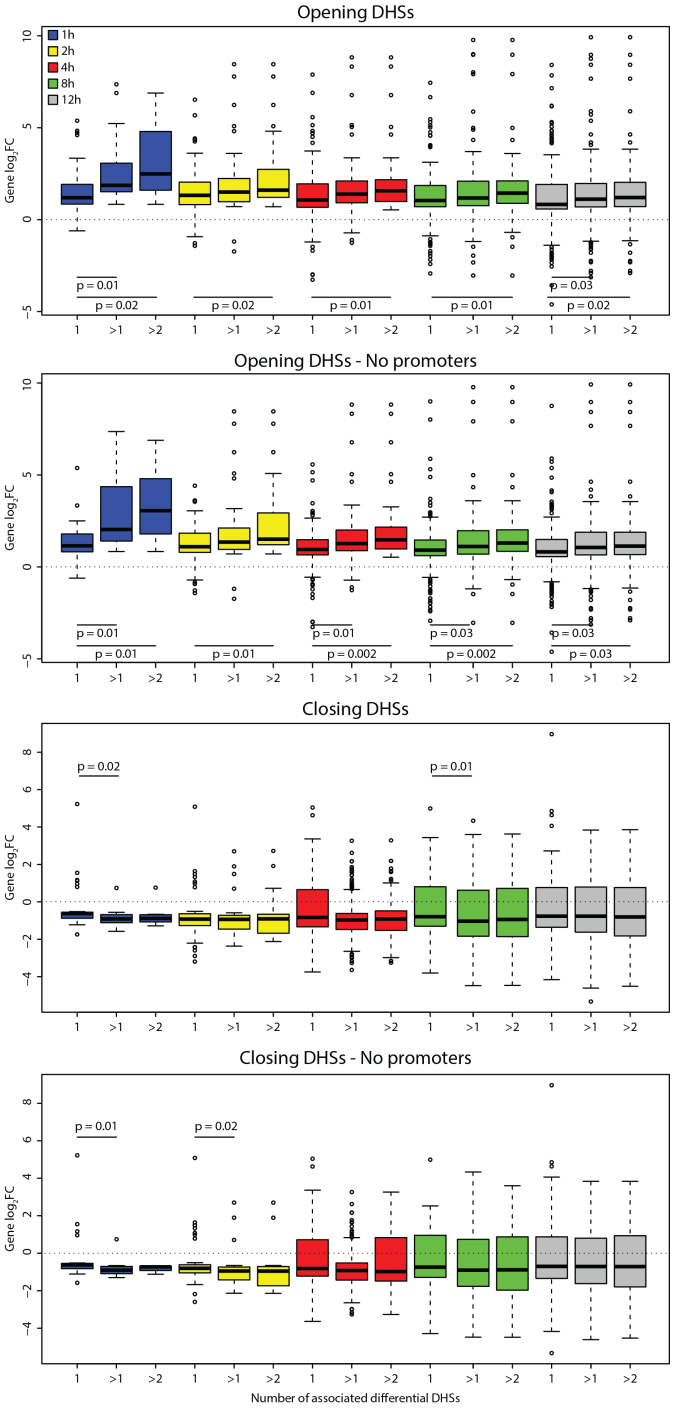

Figure 9: Distributions of $log_2(FC)$ of genes per time point, grouped by number of associated opening or closing peaks.
For each time point, genes having 1, more than 1, or more than 2 opening peaks or closing peaks were grouped together. The color of the boxplots code for the time point. The same analyses were done removing promoters from the set of differential peaks. Wilcoxon rank-sum test was used to compute p-values.

and differential peaks using hierarchical clustering, as described in 2.7. The output of the algorithm is a dendrogram that describes the sequence of merging of the clusters and the cost of each merge operation. To get the appropriate number of clusters, the dendrograms are cut where the cost of merging clusters is very high. In our case, this operation resulted in 3 clusters for accessibility dynamics (figure A.4) and 4 clusters for expression dynamics (figure A.5).

Figure 10 shows that we can define corresponding clusters between accessibility profiles and expression profiles. Generally, the observed behaviors are downregulation/closing (red cluster), upregulation/opening (white cluster) and early upregulation/early opening (green cluster). Differential genes have an additional cluster with genes that show an early downregulation followed by upregulation. A PCA of expression dynamics and accessibility dynamics shows that the profiles could be divided in 2 main clusters (figure A.6). We could have obtained the clustering observed with PCA using hierarchical clustering by cutting the dendrograms (figure A.4, A.5) at the last merge with the highest cost (data not shown).

Taken together, this suggests that response to ecdysone is generally modifying in a stable manner gene expression and chromatin landscape. Moreover, some genes and some regulatory regions have an extremely fast response, suggesting that transitory mechanisms are also present.

3.1.7 ImpulseDE2 improves modeling of the dynamics and shows similarities between accessibility and expression

The impulse model was proposed in [Chechik and Koller, 2009] to easily model with interpretable parameters the dynamics of biological systems that respond to a stimulus. To do so, the authors assume that the response follows the typical response to environmental perturbations: a first, transient adjustment that deals with immediate needs of the system, followed by a stable transition to a new steady state to adapt to the new environment. Note that it is not possible to model all types of dynamics. For instance, the impulse model is not suited to study cell cycle or circadian rythm, because they show periodical dynamics. The analysis described in 3.1.6 shows that the ecdysone response in S2 cells satisfies these assumptions.

In [Fischer et al., 2017], the impulse model was extended to be able to fit parameters using count data. Moreover, the authors presented a statistical test that is able to classify the dynamics in one of four classes:

- transition up (Tn-U) class, which represents dynamics that are monotonically increasing
- transition down (Tn-D) class, which represents dynamics that are monotonically decreasing
- transient up (Tt-U) class, which represents dynamics that are rapidly increasing followed by a decrease

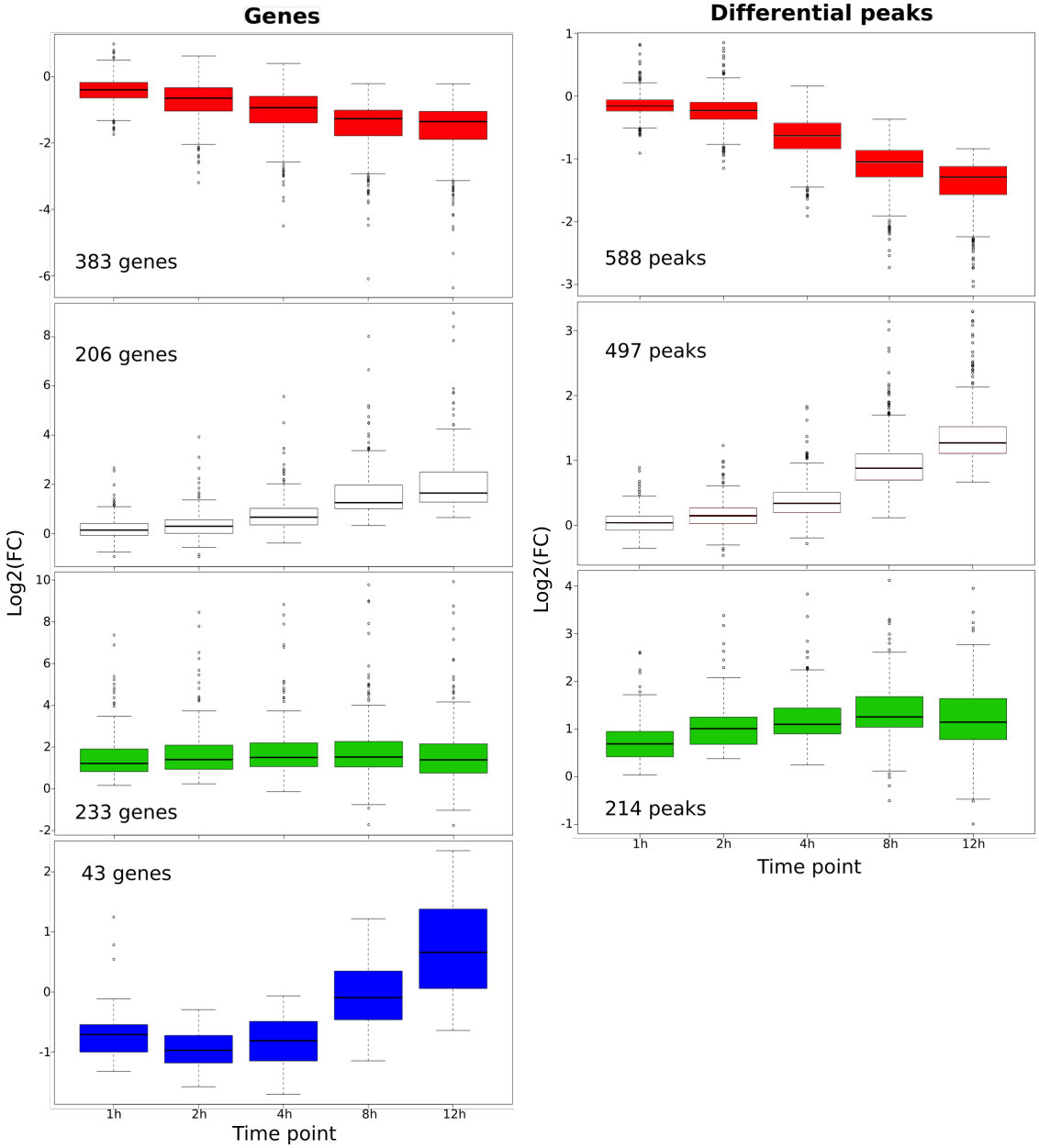

Figure 10: Distribution of $log_2(FC)$ of differential peaks and differential gene per time point, grouped by cluster.
Left: differential genes. Right: differential peaks. The color represents the behavior of each cluster. Red: downregulated genes/closing peaks. White: upregulated genes/opening peaks. Green: early upregulated genes/early opening peaks. Blue: early downregulated - late upregulated genes. In each plot the number of genes/peaks that belong to the cluster.

- transient down (Tt-D) class, which represents dynamics that are rapidly decreasing followed by an increase

The transient classes represent the immediate needs of the system after stimulation, whereas transition classes represent the new steady state.

We modeled and classified the dynamics of differential peaks and differential genes using ImpulseDE2 as respectively described in 2.4.1 and 2.5.1. ImpulseDE2 is able to detect many more differential genes and differential peaks along the time course than DESeq2 (865 differential genes and 1299 differential peaks detected by DESeq2; 1329 differential genes and 6013 differential peaks detected by ImpulseDE2). This is because with ImpulseDE2 we modeled the entire time course as a whole, whereas with DEseq2 we modeled each time point independently. The majority of profiles is classified in a *transition* class, with a percentage of 91% for both accessibility and expression (figure 11 B). This means that most of the differential genes and differential peaks are stably altered upon ecdysone stimulation, in line with the findings of [Frühauf, 2015] briefly described in 3.1.1. The relative size of the classes is very similar between differential genes classification and differential peaks classification. Moreover, standardized dynamics of differential genes and standardized dynamics of differential peaks belonging to the same class behave similarly with respect to time (figure 11 C). In the same figure we can observe that all the dynamics classified as *transient* invert their direction around 4 hours.

To validate the ImpulseDE2 classification of the dynamics of differential genes, we conducted a GO terms enrichment analysis on differential genes grouped by their class. Figure 12 shows that Tn-U genes are enriched with biological processes GO terms related to development, cell signaling and response to stimuli, whereas Tn-D genes are depleted of such terms and enriched of biological processes terms related to energy production. We also conducted a GO terms enrichment analysis on the target genes of differential peaks, grouped by class of differential peaks (figure A.7). This analysis did not give a clear separation between enriched biological processes GO terms and depleted biological processes GO terms as it did using differential genes. This was expected, because the vast majority of differential peaks do not fall on promoters (figure A.8), therefore we have a certain number of differential peaks that do not agree with the direction of the regulation of their target genes (figure A.2). Nevertheless, target genes associated with differential peaks in the Tn-U class show enrichment of terms that were enriched in the Tn-U class of differential genes (figure 12), such as terms related to development, cell signaling and response to stimuli.

3.1.8 Permanently upregulated genes show a more complicated mechanism of regulation

Given the similarity between the dynamics of differential peaks and differential genes, both in the proportion between classes and their behavior along time, and given the observation made for the GO terms enrichment analysis on the target genes of differential peaks, we quantitatively evaluated the overlap between the classes of differential genes and the target

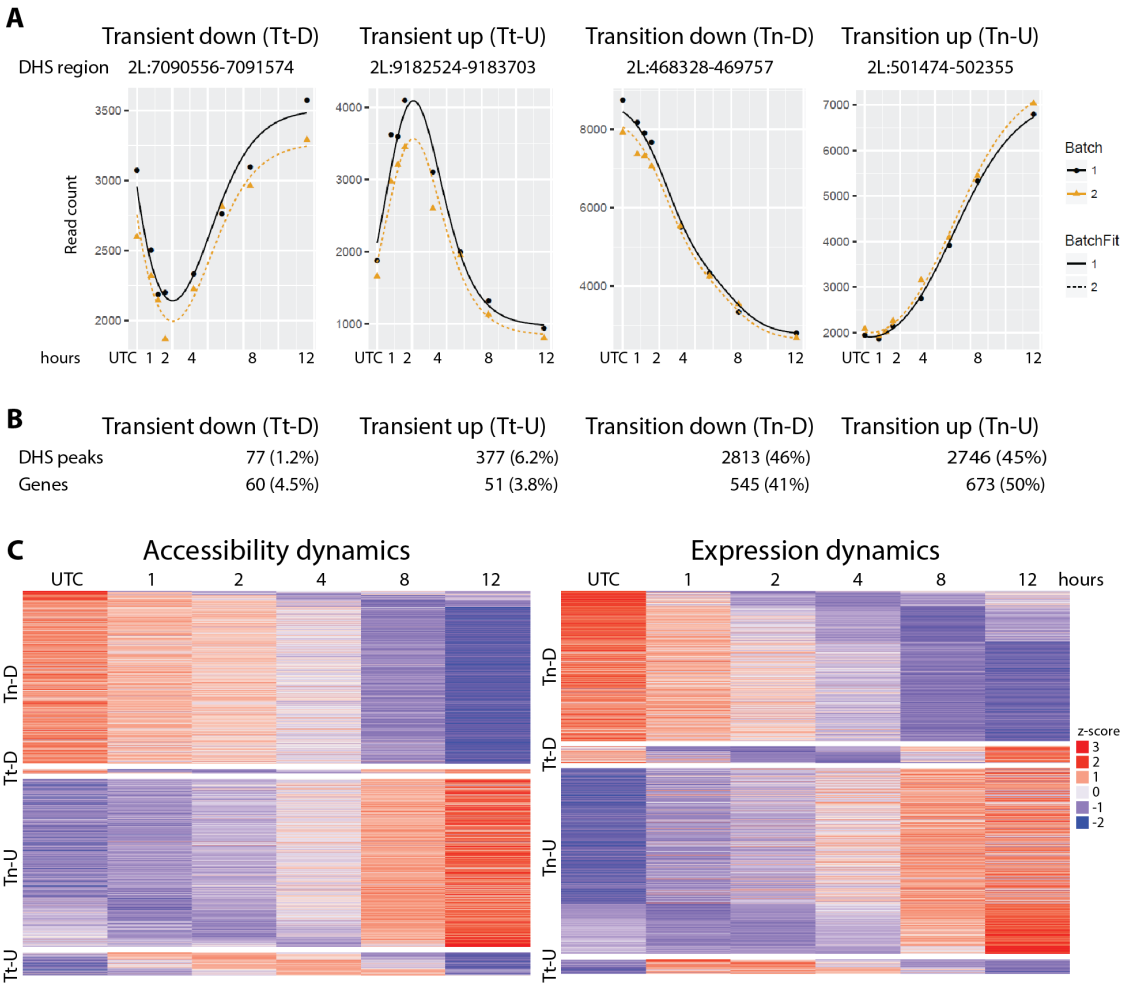

Figure 11: Modeling and classification of dynamics using ImpulseDE2.
(A) Examples of modeled profiles of 4 differential peaks, 1 per each class; (B) Number of differential genes and differential peaks belonging to each class and their percentages with respect to the total; (C) Heatmaps of accessibility dynamics and expression dynamics, grouped by class. Profiles are standardized for better visual comparison.

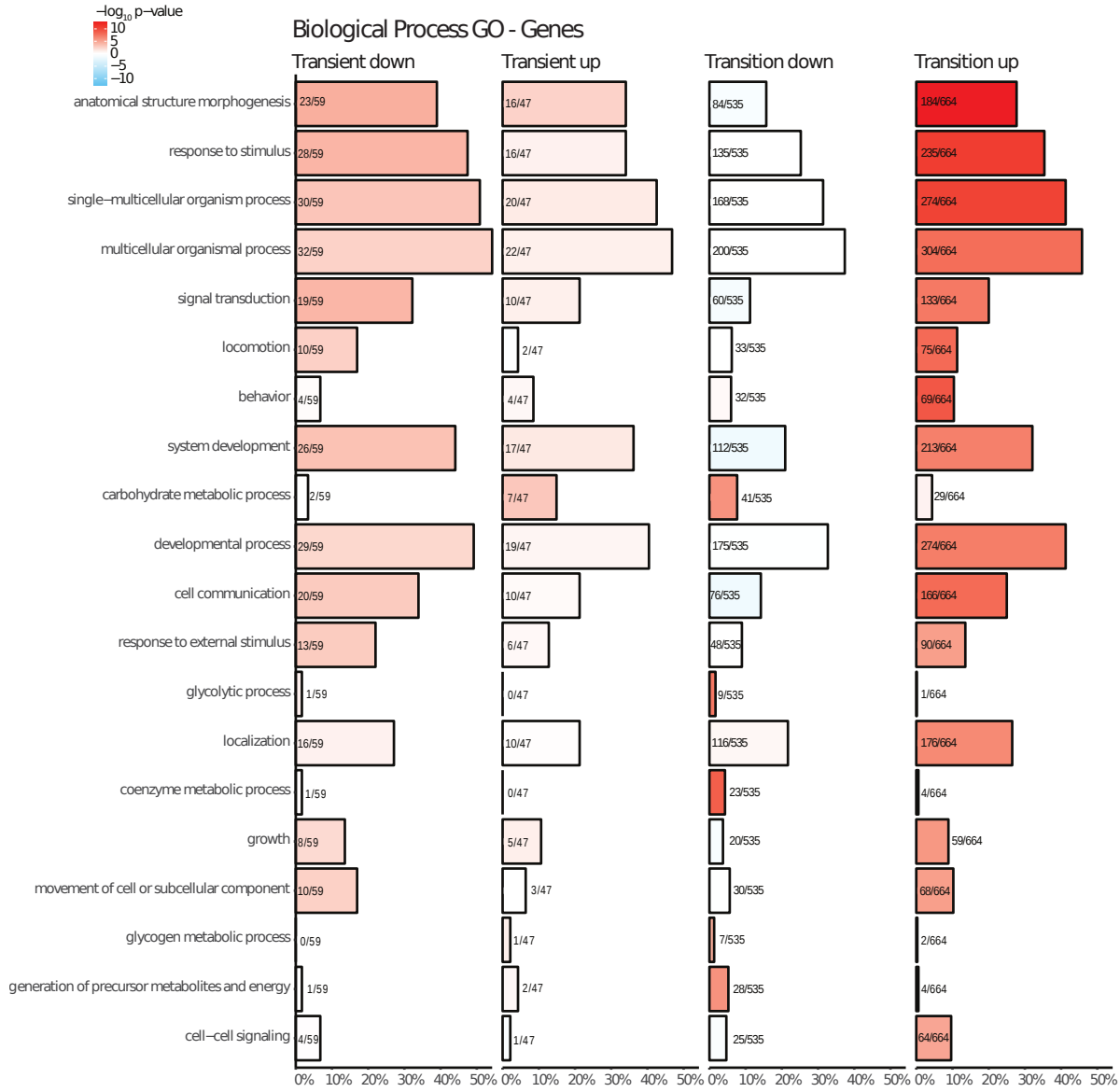

Figure 12: GO terms enrichment analysis on differential genes, grouped by class assigned from ImpulseDE2.

Bars colored in red represent enriched terms, whereas bars colored in blue represent depleted terms. Intensity of the color represents $-log_{10}(pvalues)$. Size of the bars represents the ratio (indicated on or next to the bars) between the number of genes in the class annotated to the specific term and the number of genes in the class, with the percentages scale on the x-axis.

genes grouped by classes of differential peaks. To do so we measured the Jaccard index, as described in 2.9, of all the possible combinations of sets of genes defined by the ImpulseDE2 classifications of differential genes and target genes of differential peaks.

As expected, figure 13 shows that there is agreement between the same classes and disagreement between different classes. However, differential genes classified as Tn-U, and in particular TFs, show overlap with target genes of the classes Tn-D and Tt-U of differential peaks. This suggests a general linear relationship between the openness of regulatory regions and the regulation of associated genes, with the exception of Tn-U genes that require a finer mechanism of regulation that could involve more than one layer of interactions.

3.1.9 Motif enrichment in S2 cells suggests novel TFs involved in the response

After having characterized the dynamics of differential genes and differential peaks and the relationships between them, we determined which TFs could be important for the ecdysone response in S2 cells. To do so, we selected TFs that show a differential behavior, as described in 2.11.1, and conducted a motif enrichment analysis in each class of differential peaks as described in 2.12. The full list of TFs used for the motif enrichment can be found in B.1.

Figure 14 shows the outcome of the analysis. The EcR-USP heterodimer appears enriched only in the Tt-U class, in agreement with the fact that it acts at the beginning of the ecdysone response. Br is well known to be involved in the ecdysone response, therefore its presence is not a surprise. Its enrichment in Tn-D peaks gives the most significant value for the isoform Z1, a very significant value for the isoform Z2 and it is significant in the isoform Z3.

3 TFs are enriched in all classes of differential peaks: hng3, srp and the already mentioned br-Z1. Hng3 belongs to the MADF-BESS domain transcription regulators group, which includes chromatin modifying proteins, and it is not known to be associated with the ecdysone response. Srp has been recently suggested to be required in enhancers that are ecdysone-induced in S2 cells [Shlyueva et al., 2014]. In agreement with this result, in our data it is enriched at most in the Tn-U class of differential peaks.

3 further TFs that are highly enriched are CG5953, pnr and foxo. CG5953 belongs to the MADF-BESS domain transcription regulators group and it is not known to be involved in ecdysone response. Pnr plays a role in the development of imaginal discs and nervous system and it is not known to be involved in the ecdysone response. Foxo is involved in the regulation of the insulin signaling pathway and it has been related to ecdysone [Koyama et al., 2014].

Overall, our motif enrichment analysis recovers TFs that have already been associated to the ecdysone response, and it suggests new ones.

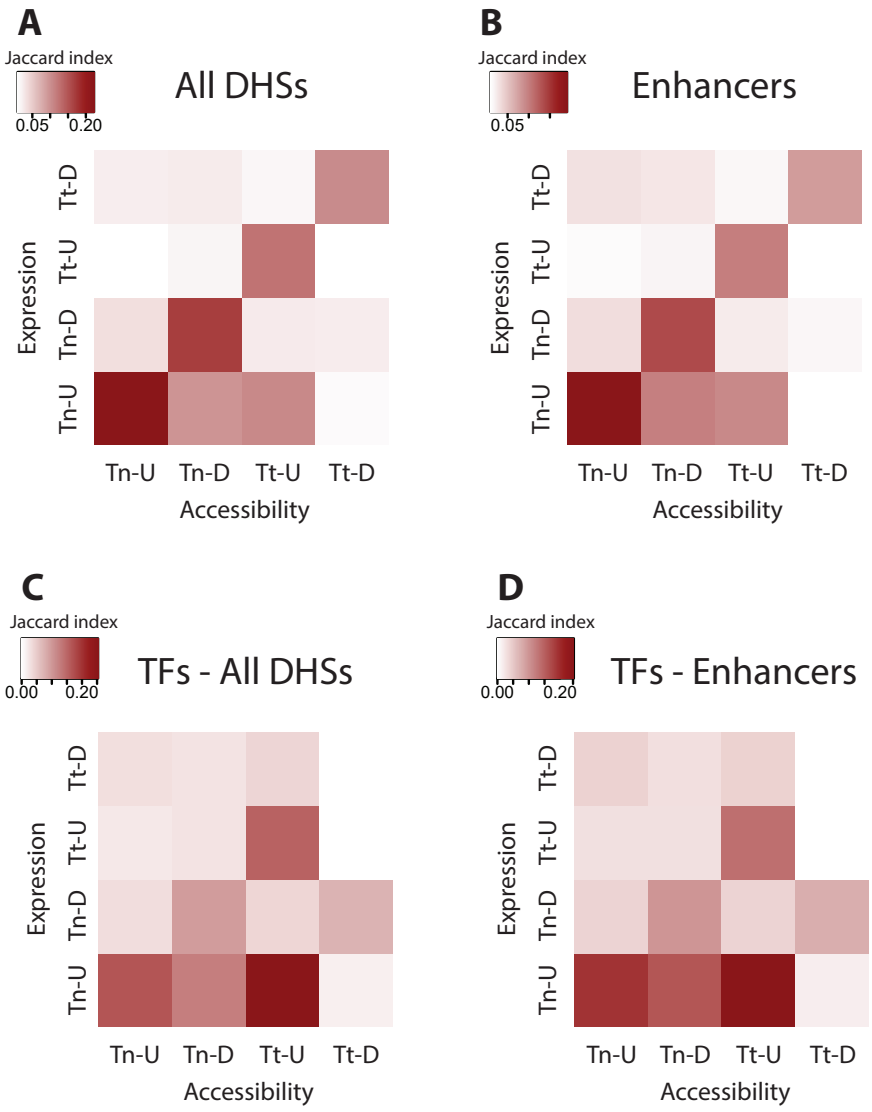

Figure 13: Similarities between sets of differential genes and target genes of differential peaks, grouped by class.
X-axis: sets of target genes, defined by class of associated differential peaks. Y-axis: sets of differential genes, defined by their class. The intensity of the color reflects the degree of overlap. (A) Similarity measured using all differential peaks and all differential genes. (B) Similarity measured using only enhancers differential peaks and all differential genes. (C) Similarity measured using all differential peaks and differential genes that are TFs. (D) Similarity measured using only enhancers differential peaks and differential genes that are TFs.

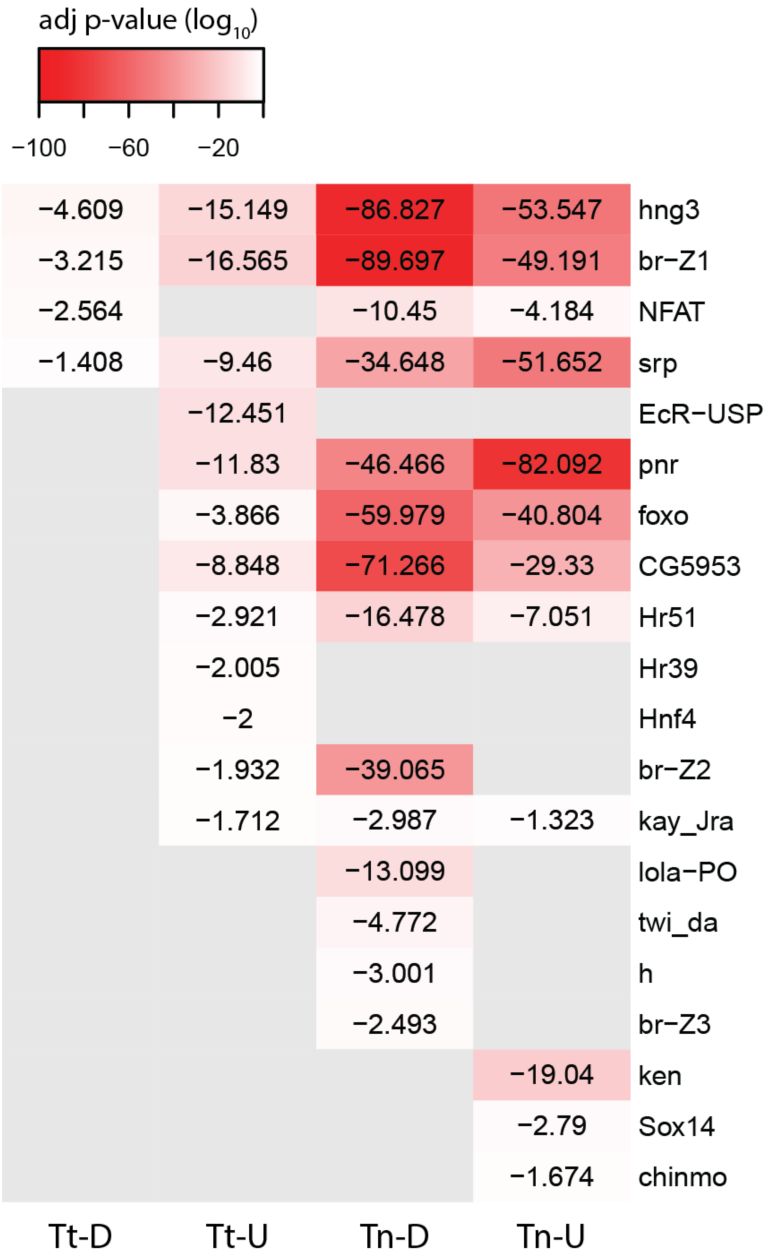

Figure 14: Enrichment of motifs of TFs with differential behavior in S2 cells, computed for each class of differential peaks.
X-axis: classes of differential peaks. Y-axis: TFs enriched in at least one class of differential peaks. The $log_{10}(pvalue)$ of each enrichment is reported. The intensity of the color represents the significance of the enrichment.

3.2 Characterization of the chromatin landscapes during pupariation

3.2.1 ATAC-seq reliably captures chromatin landscapes across tissues

As described in 1.3, the ecdysone response in larvae has extreme effects, which range from survival and differentiation of imaginal tissues to programmed cell death of SG. These different effects are all triggered from the ecdysone-ligated TF EcR. Subsequently, different cofactors are involved in the tissue-specific responses, precisely determining cell fates and making it an interesting paradigm to study.

To characterize chromatin landscapes during pupariation, we selected 3 stages that encompass the ecdysone pulse responsible for pupariation (E3IL, L3IL and WPP) and 4 tissues that represent the entire range of responses to the ecdysone pulse (ED, WD, SG and CNS) and gathered ATAC-seq data to have a picture of the chromatin landscape during larval development (figure 15 B), as described in 2.1. ATAC-seq was chosen because it has a very fast protocol and requires small amounts of starting material, as described in 1.2.3, making it suitable to be used in this paradigm.

To assess the quality of our ATAC-seq data, we checked how they correlated with our DNase-seq data on the same sample. As expected, we found a very high correlation both genome-wide and in detected peaks (figure A.9), so we are confident to identify functional regulatory regions.

A qualitative analysis at the EcR locus shows that our data capture the chromatin changes along time and across tissues. Chromatin landscapes among different tissues at the same stage are more different than chromatin landscapes in the same tissue across stages (figure 15 B), in agreement with the findings of [McKay and Lieb, 2013]. Moreover, WD and ED have a similar chromatin landscape, in line with their common fate. By contrast, CNS and SG show different sets of active regulatory regions, reflecting the fact that they undergo different fates. Since our data is able to capture the different chromatin landscapes, we analyzed quantitatively the differences between them.

3.2.2 Chromatin landscapes reflect tissues fates

To measure the similarities between chromatin landscapes of different tissues, we measured the distances between samples as described in 2.10. As observed qualitatively, CNS and SG are distant between each other and also from WD and ED (figure 16 A, B and C), in line with the fact that SG undergo programmed cell death whereas CNS has a mixed fate. By contrast, WD and ED chromatin landscapes are very similar along all stages, reflecting their shared fates of differentiation. This is also confirmed by detecting differential peaks across tissues in the same stage, as described in 2.4.2. The number of differential peaks between WD and ED in all stages is substantially lower than the rest of the comparison (data not shown). As expected, after quantification chromatin landscapes among different

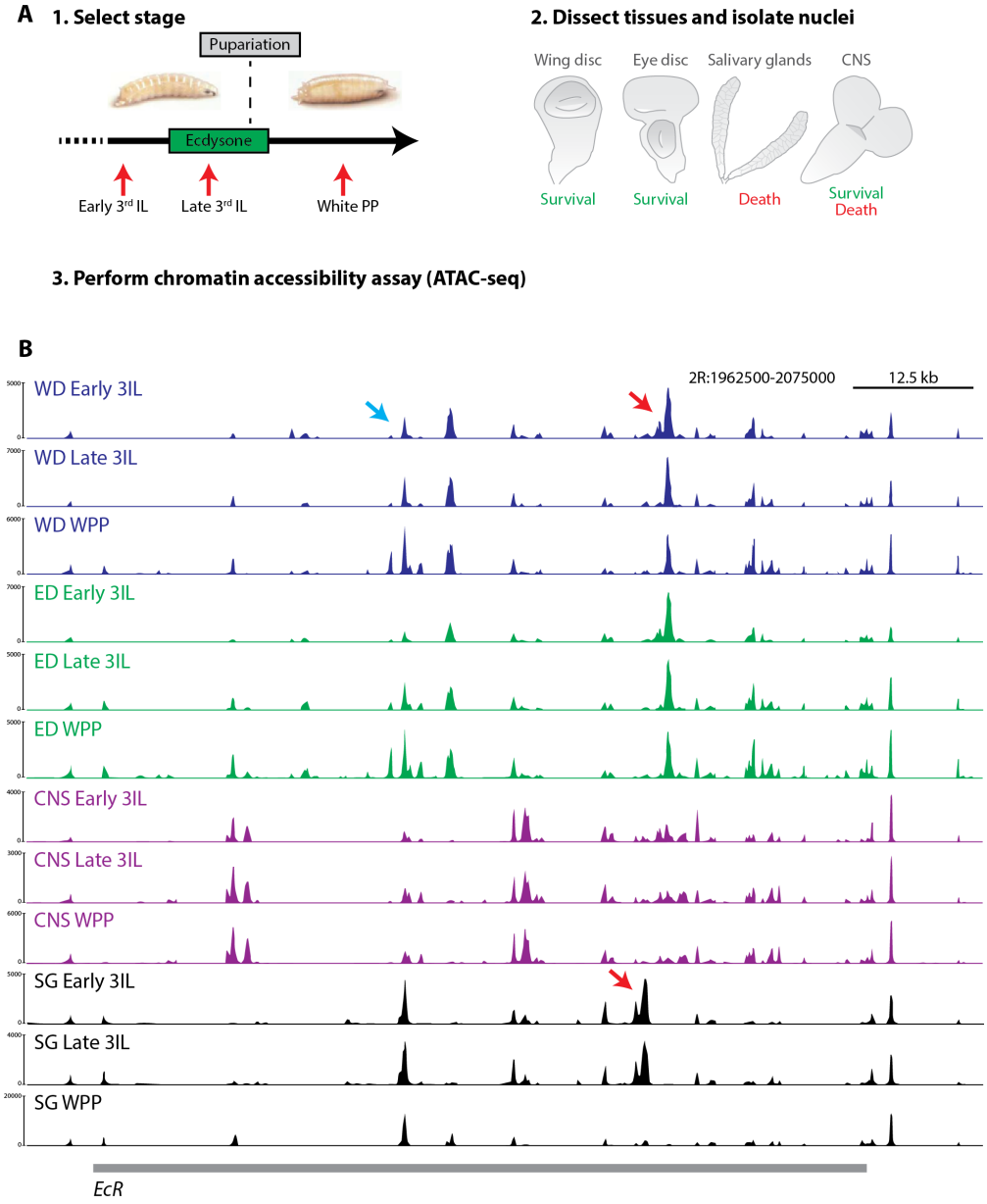

Figure 15: Experimental setup of larval paradigm and chromatin landscape at the EcR locus.
(A) Visualization of the experimental setup. The selection of the stages during the ecdysone pulse and the fates of the chosen tissues are depicted. (B) Chromatin landscape of every stage and every tissue at the EcR locus. Violet: WD. Green: ED. Purple: CNS. Black: SG. The main differences between tissues are highlighted with arrows.

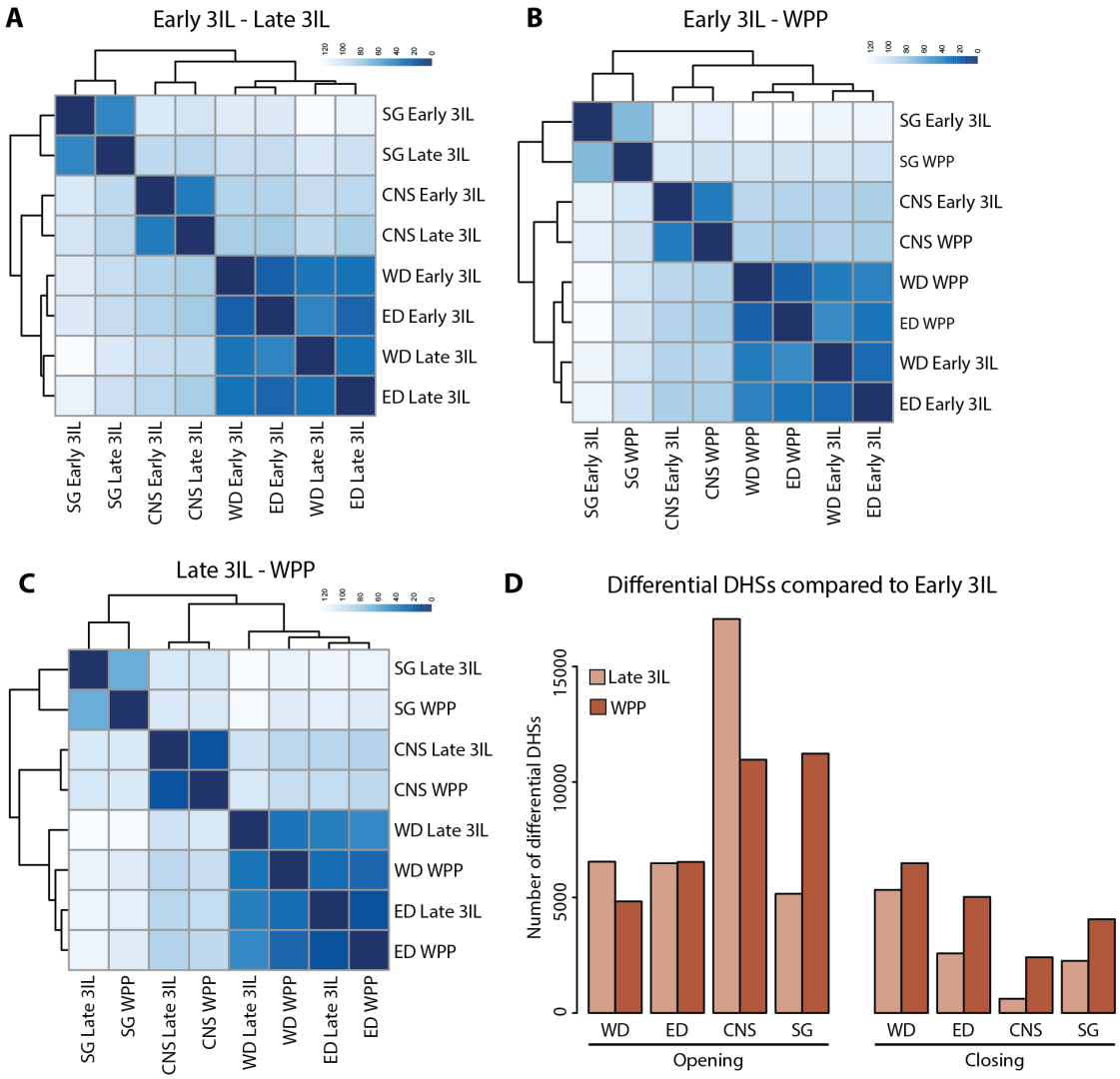

Figure 16: Differences in chromatin landscapes of larval tissues are measured using distances between samples.
(A) Distances measured between E3IL and L3IL stages. (B) Distances measured between E3IL and WPP stages. (C) Distances measured between L3IL and WPP stages. The intensity of the blue reflects the closeness of 2 samples. (D) Number of differential peaks in L3IL and WPP compared with E3IL, grouped by tissue and separated by direction. Light brown: L3IL. Dark brown: WPP.

tissues at the same stage are more distant than chromatin landscapes in the same tissue across stages.

We called differential peaks along stages as described in 2.4.2, using E3IL as reference stage. The number of opening peaks in L3IL CNS is very high compared with the other tissues in the same stage, whereas the number of closing peaks is very low (figure 16 D).

Taken together, this suggests that the fates of the different tissues are reflected in their chromatin landscapes. Moreover, if tissues have a similar cell fate, the dynamics of their differential peaks are also similar.

3.2.3 Tissue-specific motif enrichment suggests TFs involved in the response

After having described the relationship between chromatin landscapes and cell fates, we determined which TFs could be important for the ecdysone response in larval tissues. To do so, for each tissue we selected TFs that show differential behavior, as described in 2.11.2, and conducted a motif enrichment analysis for each type of differential peak in each stage, as described in 2.12. Since we did not have expression data, we defined differential behavior as having a differential peaks along stages on the promoter. This is justified because there is a correlation between $log_2(FC)$ of accessibility at promoters and expression in S2 cells (figure A.3). The full lists of TFs used for the motif enrichments can be found in B.2 for WD, B.3 for ED, B.4 for CNS and B.5 for SG. The full lists of enrichments can be found in B.6 for WD, B.7 for ED, B.8 for CNS and B.9 for SG. To be tissue-specific, we removed all TFs that were present in at least 2 tissues, leaving only enriched motifs present in exactly one tissue.

Not surprisingly, ED and WD show only a few enriched TFs (figure 17 and 18). This is because the enriched motifs are redundant between the 2 tissues, as they share the fate of survival and differentiation. In WD br-Z2 is enriched for both time points in closing peaks, suggesting that it plays a role in chromatin repression. In SG bin is known to play a role in salivary gland morphogenesis. Indeed, it is the most enriched motif in all the 4 enrichment for SG (figure 20). Bab1 is enriched in all stages and all types of differential peaks in CNS. Drosophila has 2600 olfactory receptor neurons [Li and Liberles, 2015], and bab1 is involved in their fate diversity. Again in CNS, crc is enriched in opening peaks in the L3IL stage. Crc is an EcR co-activator, therefore its enrichment is in line with its known role. HGTX, enriched in opening peaks in the CNS WPP stage, is known to promote development and differentiation of motor neurons, and it is known to have a major role in neuronal specification and differentiation. In general, the complexity of the enrichment of CNS, in particular the L3IL stage, is in line with the complexity of the response that the tissue has to ecdysone, partially differentiating and partially dying (figure 19).

Overall, our motif enrichment analysis recovers TFs that have already been associate to the ecdysone response in the different tissues, and it suggests new ones.

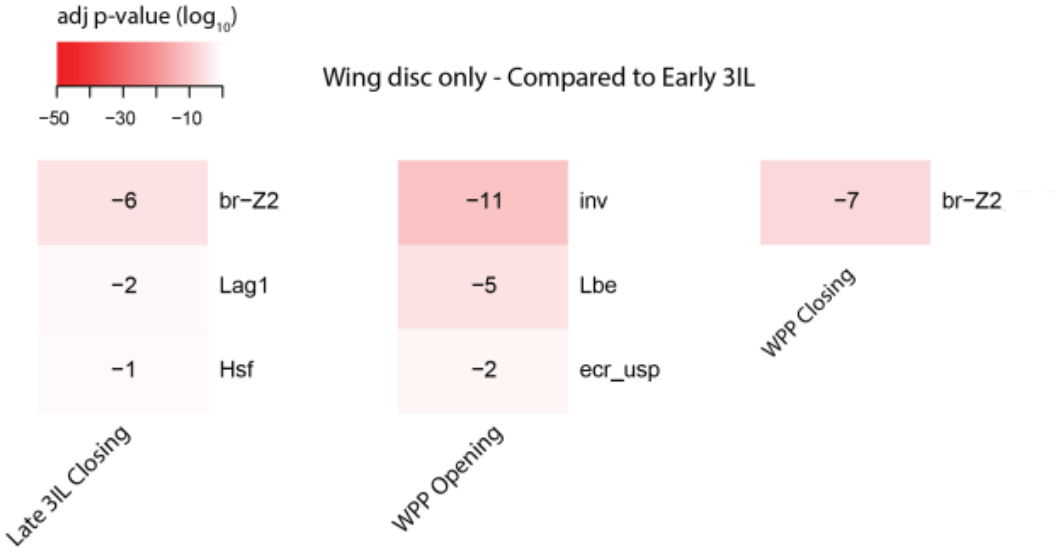

Figure 17: Enrichment of motifs of TFs with differential behavior in WD, computed for each type of differential peak in each stage.
X-axis: type and stage of differential peaks. Y-axis: TFs enriched in at least one type and stage of differential peaks. Motifs that were enriched in ED, CNS or SG were removed. The $log_{10}(pvalue)$ of each enrichment is reported. The intensity of the color represents the significance of the enrichment.

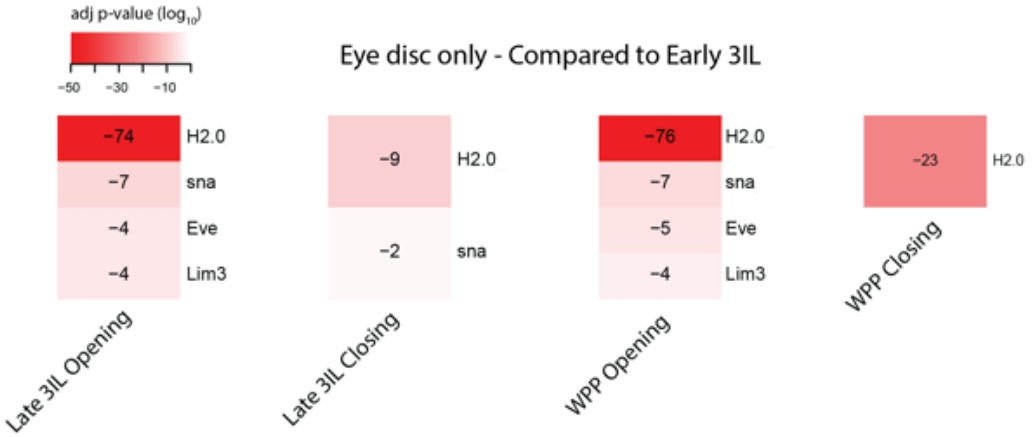

Figure 18: Enrichment of motifs of TFs with differential behavior in ED, computed for each type of differential peak in each stage.
X-axis: type and stage of differential peaks. Y-axis: TFs enriched in at least one type and stage of differential peaks. Motifs that were enriched in WD, CNS or SG were removed. The $log_{10}(pvalue)$ of each enrichment is reported. The intensity of the color represents the significance of the enrichment.

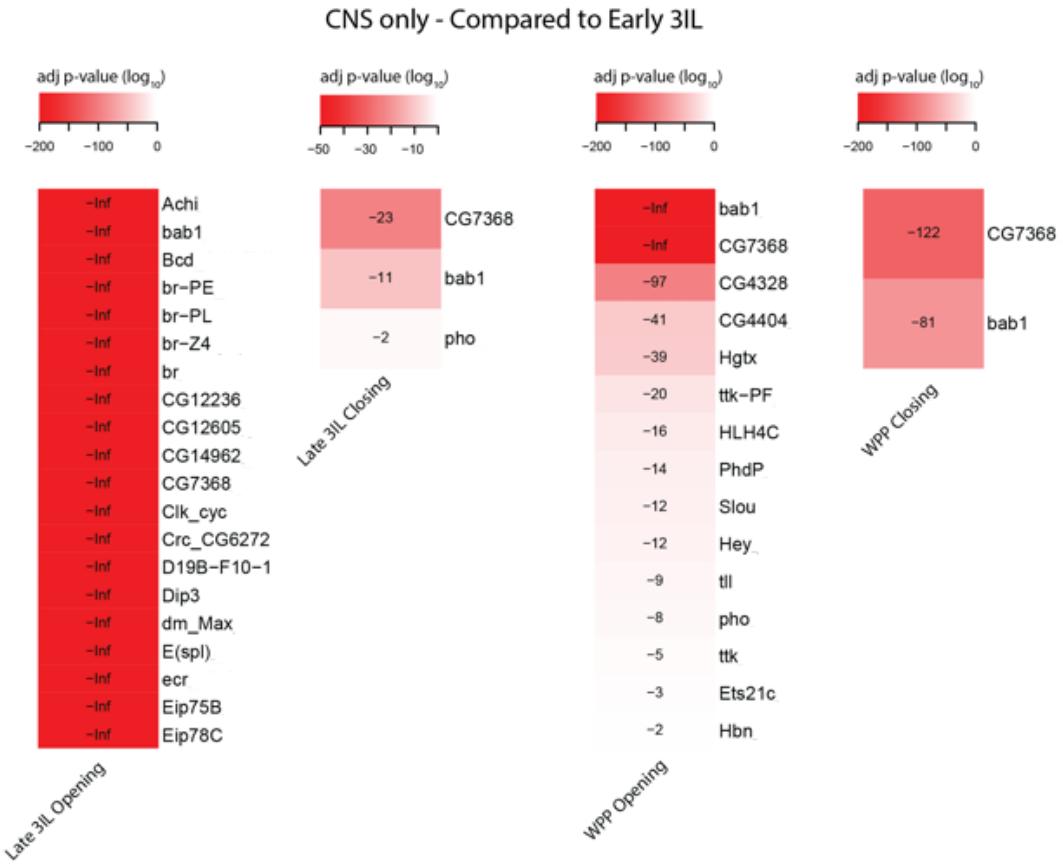

Figure 19: Enrichment of motifs of TFs with differential behavior in CNS, computed for each type of differential peak in each stage.
X-axis: type and stage of differential peaks. Y-axis: TFs enriched in at least one type and stage of differential peaks. Motifs that were enriched in WD, ED or SG were removed. The $log_{10}(pvalue)$ of each enrichment is reported. The intensity of the color represents the significance of the enrichment.

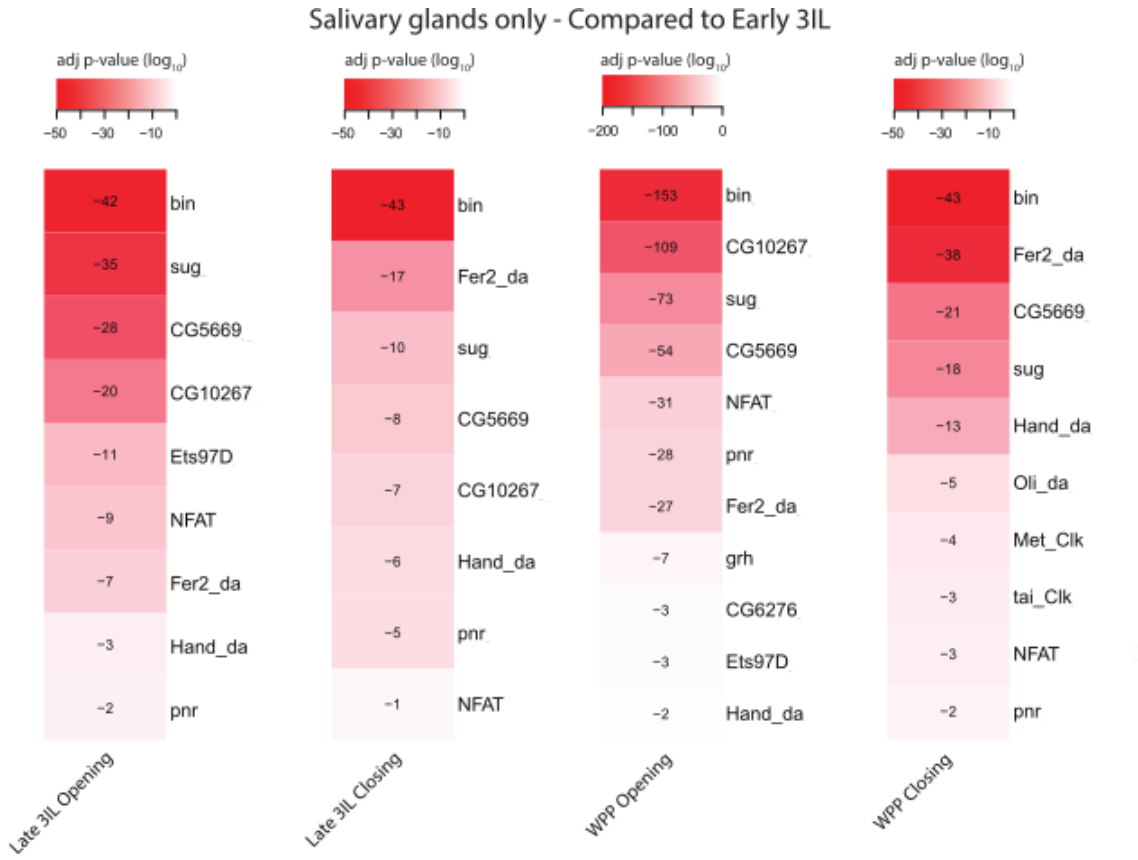

Figure 20: Enrichment of motifs of TFs with differential behavior in SG, computed for each type of differential peak in each stage.
X-axis: type and stage of differential peaks. Y-axis: TFs enriched in at least one type and stage of differential peaks. Motifs that were enriched in WD, ED or CNS were removed. The $log_{10}(pvalue)$ of each enrichment is reported. The intensity of the color represents the significance of the enrichment.

3.3 Statistical modeling of the ecdysone response in S2 cells

So far, we characterized the dynamics of differential peaks and differential genes in S2 cells, and suggested new TFs that could be involved in the ecdysone response using motif enrichment. Can we do more? Can we model gene expression using accessibility data and TFs affinities in a way that is easy to interpret, and that gives us additional information on TFs involved in the ecdysone response and their function?

The answer is yes, and we did so using regularized linear regression and regularized logistic regression as described in [Schmidt et al., 2016, Durek et al., 2016]. We used these models because of their easiness in the interpretation of the results, and because they use a TFs affinities prediction method called TRAP [Roider et al., 2006], which is able to include TFBSs with low affinity in the calculation. As we have already mentioned in 1.1.3, weak binding plays an important role in regulation of gene expression [Tanay, 2006, Segal et al., 2008].

We tested two novel elements, and successfully introduced one of them, in the calculation of TF-gene scores used as features by the models: the expression level of TFs, to decrease or increase the TF-gene score according to the expression of the TF taken into consideration, and the method to assign target genes to peaks, to see if we could fit better models.

3.3.1 Definition of the independent variables

The first operation is to define a comprehensive set of TFs that could be involved in the ecdysone response cascade. In contrast to the strict selection that we used to define the set used for the motif enrichment in 3.1.9, here we define the set in a permissive way, as described in 2.11.3, in order to not miss any TF, and let the models tell us which are not important. The list of TFs can be found in B.10.

The next step is the calculation of TF-gene scores, which are used as independent variables, as described in 2.13. TF-gene scores are computed for each pair of differential gene and TF, in all the time points measured in the ecdysone response in S2 cells. We tested all the possible combinations of factors in the TF-gene score definition: with or without multiplying by the expression level r_j of TF j, with or without multiplying by the mean accessibility s_p of peak p and with or without multiplying by the exponential decay $e^{-\frac{d_{p,i}}{d_0}}$ given by the distance between peak p and TSS of gene i.

Including the expression level r_j of TF j modifies the TF-gene scores to reflect the abundance of j in the system. In fact, if TF j is not expressed in the system, or it has a very low expression level, it can not bind its TFBSs and it can not regulate any gene. Therefore, when the expression level r_j is very low, also the TF-gene scores will be very low. By contrast, if TF j is highly expressed in the system, the TF-gene scores will be very high to reflect its abundance.

The mean accessibility s_p of peak p modifies the TF-gene scores to reflect the propensity

of peak p to be bound by TFs. If peak p is open, it will be more easily bound by TFs, which will regulate the target genes. As a consequence, a high mean accessibility s_p increases the TF-gene scores. By contrast, if peak p is closed, typically it can not be bound by TFs, and a low mean accessibility s_p will decrease the TF-gene scores.

The exponential decay decreases the contribution of each peak with the distance from the TSS. If it is used, it is assumed that distal regulatory regions have a weaker influence on gene regulation. Conversely, if it is not used, it is assumed that distal regulatory regions have the same influence on gene regulation as the proximal regulatory regions.

Definition of the set P_i of peaks assigned to gene i

As already mentioned in 3.1.3, the assignment of target genes to enhancers is not a trivial problem. We tested three different strategies to define the set P_i of peaks assigned to gene i: assignment by nearest TSS, assignment by the definition of regions of influences of genes and assignment by windows centered on the TSSs. The three strategies are defined in 2.6 and a visualization of the assignments is shown in figure 21, where the strategies are depicted in terms of the regions in which they segment the genome. For each strategy, peaks are assigned to the genes whose regions belong to.

The three strategies are different between each other. The regions of *nearest TSS* do not overlap between each other. This means that each peak is assigned to exactly one gene. Moreover, the regions are defined only by the distances between TSSs. If the distance between TSSs is high, the region will be long, whereas if the distance between TSSs is low, the region will be short. *Window on TSS* and *region of influence* strategies allow for peaks to be assigned to multiple genes. However, regions are defined in different ways. *Region of influence* regions span at least the entire gene body, whereas *window on TSS* regions are centered on TSSs of genes and have a fixed size.

As we already mentioned in 3.1.3, in Drosophila the regulatory regions of a gene are not very distant from the gene itself. In the assignment by definition of regions of influences, there are 2 parameters D and α that regulate the extension of the region if a gene has distant neighbors. In particular, D regulates the minimum distance for which 2 genes are considered distant, whereas α regulates how much the region of influence of a gene gets close to the neighbors. In our tests, we used $D = 20000$ and $\alpha = \frac{1}{2}$.

We tested the following widths w for the assignment by windows centered on TSSs: 1k, 2k, 5k, 10k, 20k and 50k bp.

3.3.2 Regularized linear regression suggests functionalities of TFs in the ecdysone response

To suggest the functionalities of TFs, we fit a regularized linear regression as described in 2.14, using cross validation to avoid overfitting as described in 2.16. We tested different TF-gene scores definitions and different assignments of peaks to target genes, as described in 3.3.1. To do so, for all the possible combinations of factors and assignments of target genes we calculated TF-gene scores, and we fit a model for each time point. We chose to

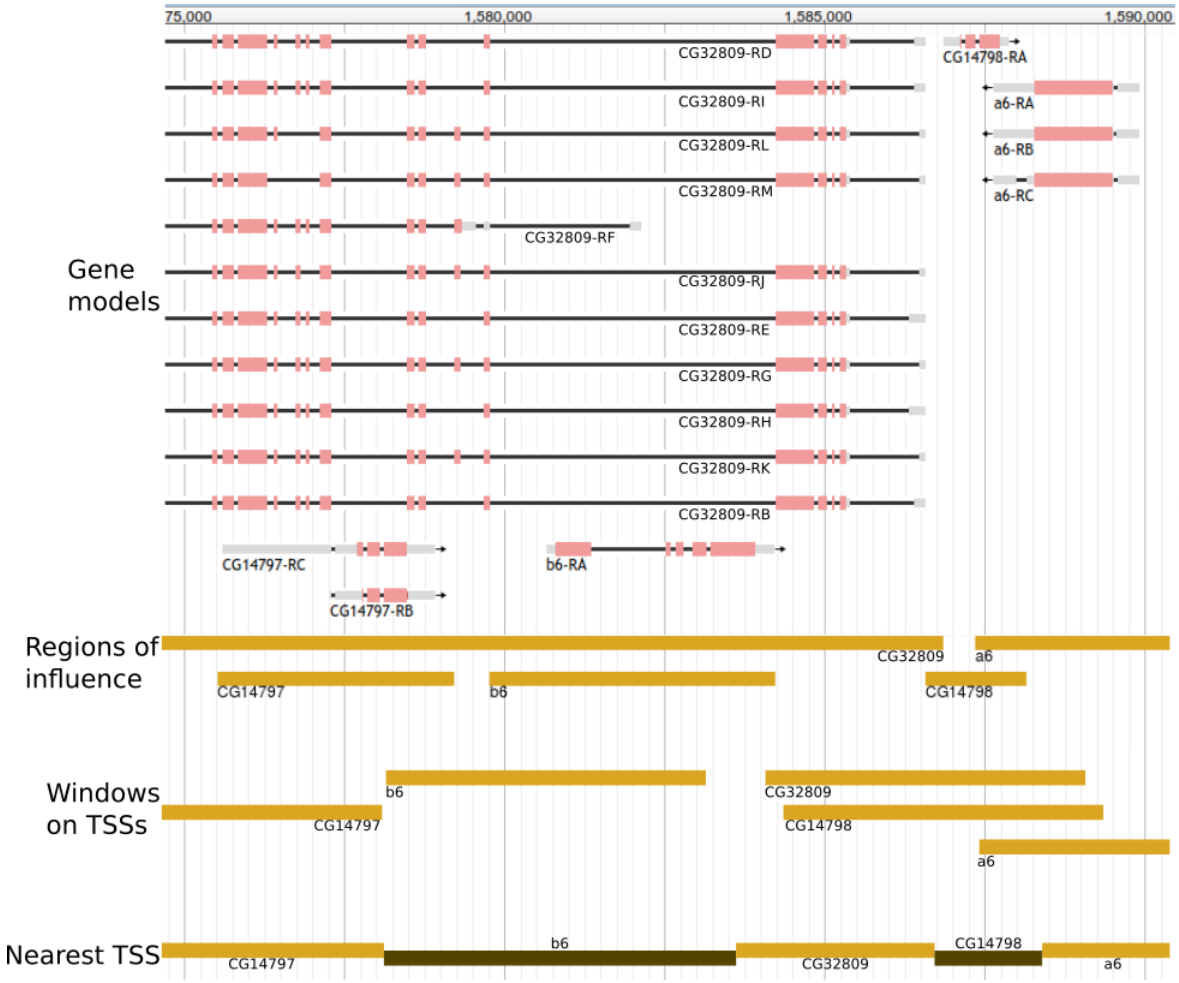

Figure 21: Visualization of the strategies to assign target genes to peaks. Screenshot of the genome browser with the regions defined by each strategy. First row: gene models. Second row: regions defined by the *regions of influence* strategy. Third row: regions defined by the *window on TSS* strategy, 5k bp windows. Fourth row: regions defined by the *nearest TSS* strategy. Peaks that overlap a region are assigned to the gene whose region belongs to.

use the TF-gene scores that gave the lowest average mean squared error (MSE) between measured gene expression and predicted gene expression along the time course. For each time point, the mean MSE was calculated averaging the MSE of the test sets of the outer cross validation. A full list of performances is reported in B.11. An example of scatterplots of a time course between measured gene expression and predicted gene expression is reported in figure 22. The corresponding average PCC along the time course between measured gene expression and predicted gene expression is 0.396, and we obtained it with the following setup:

- $x_{i,j} = r_j \sum_{p \in P_i} a_{p,j} e^{-\frac{d_{p,i}}{d_0}}$
- P_i is defined using the assignment based on window centered on TSSs, $w = 50k$ bp

This means that TF-gene scores computed using the expression of TFs, without the mean accessibility, with the exponential downweighting and with the window on TSS strategy, with a window of 50k bp, gave the best performances. A qualitative analysis of the results with the 20 best performances shows that the exponential downweight is a factor that helps the fitting, together with the assignment with the window on TSS strategy. The expression of TFs generally improves the performances, however there are results with good performances that don't use it. Not factoring the mean accessibility into the TF-gene scores generally gives the very best performances. However, when we take a broader look, results for this factor are more inconsistent. Hypothesis on how these factors influence performances will be discussed in 4.3.

The regularization that we use is the elastic net [Zou and Hastie, 2005], which is a mixture of L1-norm regularization, also called LASSO, and L2-norm regularization, also called Tikhonov regularization or ridge regression. Ridge regression keeps the estimated regression coefficients $\hat{\beta}$ small, but it is not able to set them to 0. By contrast, LASSO has the desirable property that it is able to set estimated regression coefficients $\hat{\beta}$ to zero, allowing for an easier interpretation of the features of the model. However, in the case of groups of independent variables with high pairwise correlations, LASSO tends to select one independent variable for each group and it does not have a preference for which one. The elastic net overcomes this problem, because it selects groups of correlated independent variables and distributes the weights among them [Zou and Hastie, 2005]. Groups of independent variables with high pairwise correlations are not to be overlooked, because they corresponds to TFs that act together to co-regulate gene expression.

The interpretation of the regression coefficients is conceptually straightforward. If the estimated coefficient for TF j is positive, it means that if we increase the TF score for TF j the expression levels increase. Viceversa, if the estimated coefficient is negative, if we increase the TF score the expression levels decrease. In other words, a positive estimated coefficient suggests that a TF is an activator, whereas a negative estimated coefficient suggests that a TF is a repressor.

Figure 23 shows the estimated regression coefficients for the fit with the best performances. The first interesting observation is that EcR-USP gets a negative coefficient at UTC. It has been reported that in the margin of wing discs EcR-USP act as a repressor

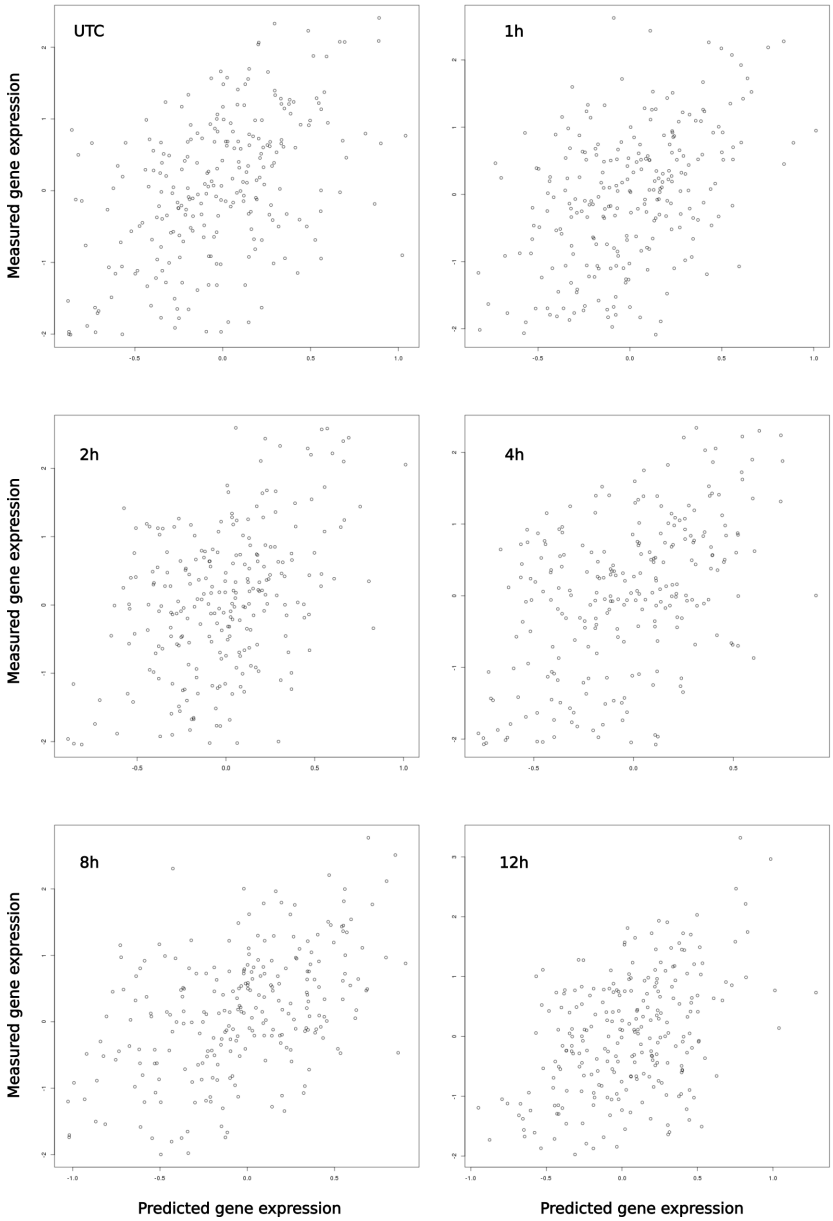

Figure 22: Example of scatterplots of a time course between measured and predicted gene expression.
X-axis: predicted gene expression. Y-axis: measured gene expression. UTC PCC: 0.473. 1h PCC: 0.447. 2h PCC: 0.428. 4h PCC: 0.484. 8h PCC: 0.473. 12h PCC: 0.437. The scatterplots represent the test sets of the cross validation with the highest performances, with the TF-gene scores that gave the lowest mean MSE along the time course.

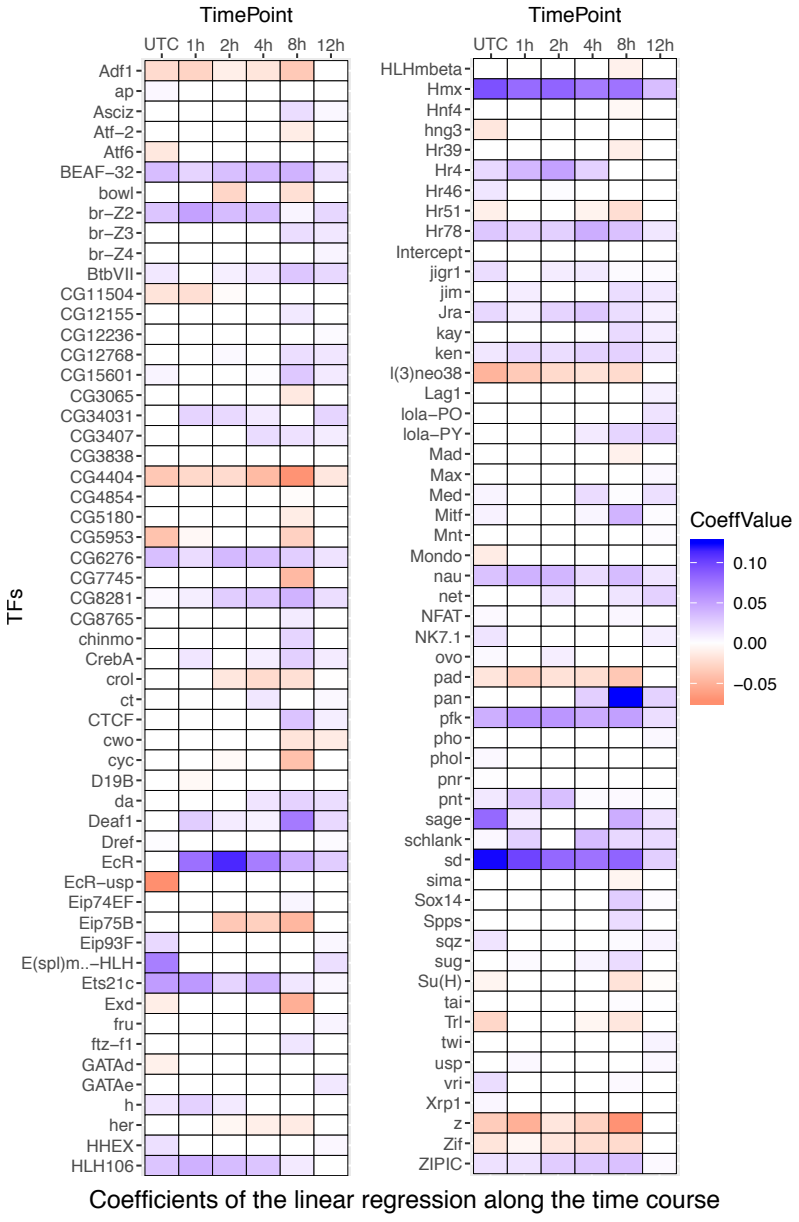

Figure 23: Estimated linear regression coefficients along the time course for each TF.

X-axis: time points. Y-axis: TFs with at least one estimated coefficient different from 0. The blue color denotes a positive estimated coefficient, which suggests an activating role, whereas the red color denotes a negative estimated coefficient, which suggests a repressing role.

if ecdysone has not ligated [Schubiger et al., 2005]. This seems to hold also in S2 cells. After stimulation, EcR is suggested to be an activator, in particular in early time points, whereas in later time points its coefficient decreases. This decreasement could indicate that EcR starts to have a repressor effect, possibly in an indirect fashion. Eip75B is suggested to act as a repressor. This has been reported in the literature [Hiruma and Riddiford, 2004, Horner et al., 1995, Reinking et al., 2005, Sullivan and Thummel, 2003, White et al., 1997]. Moreover, it has been reported to suppress Hr51 [Rabinovich et al., 2016]. Sage is reported to be an activator, but not to be involved in the ecdysone response. Generally, estimated coefficients are consistend along the time course, increasing our confidence in the suggestions. To check the agreement between suggested TFs functionalities and functionalities reported in the literature, we developed a test based on GO terms with experimental evidence. Our suggestions agree with roughly half of the published functionalities.

3.3.3 Ratio of TF-gene scores represents variations of TFs impact

To represent the variation of score between TFs and genes along the time course, we used the ratios between TF-gene scores of different time points. These ratios will be used as independent variables in a regularized logistic regression model to suggest which TFs are important to explain the observed differential expression. For a particular gene and a particular TF, if the ratio is bigger than 1, it means that the TF-gene score is higher in later time points, and such TF has a bigger impact in the regulation of such gene. By contrast, if the ratio is lower than 1, it means that the TF-gene score is lower in later time points, and such TF has a smaller impact in the regulation of such gene. To validate TF-gene score ratios, we visualized them in heatmaps for each differential gene and for each TF. A sample is reported in figure 24, which depicts log_2 of TF-gene score ratios for Eip93F and Eip78C. We can observe positive values of log_2 of TF-gene score ratios for EcR-USP. This is expected, since Eip93F and Eip78C are early ecdysone-responsive genes. Moreover, we can observe positive values for br, which likely regulates them. In general, TF-gene score ratios can be used to determine which TFs may have an impact on the regulation of each single gene. Given this validation, we assume that TF-gene score ratios correctly represent variation of impact of TFs on gene regulation.

3.3.4 Regularized logistic regression suggests TFs responsible for differential expression in the ecdysone response

To suggest which TFs are responsible for gene regulation in the ecdysone response, we fit a regularized logistic regression as described in 2.15, using cross validation to avoid overfitting as described in 2.16. We chose the same elastic net regularization as for linear regression. We always compared each time point after stimulation with UTC, and fit a model for each comparison. Logistic regression is used when the dependent variable has a binary outcome, in our case a gene being upregulated or downregulated, and one wants to estimate the probability of the outcome based on independent variables, in our case the

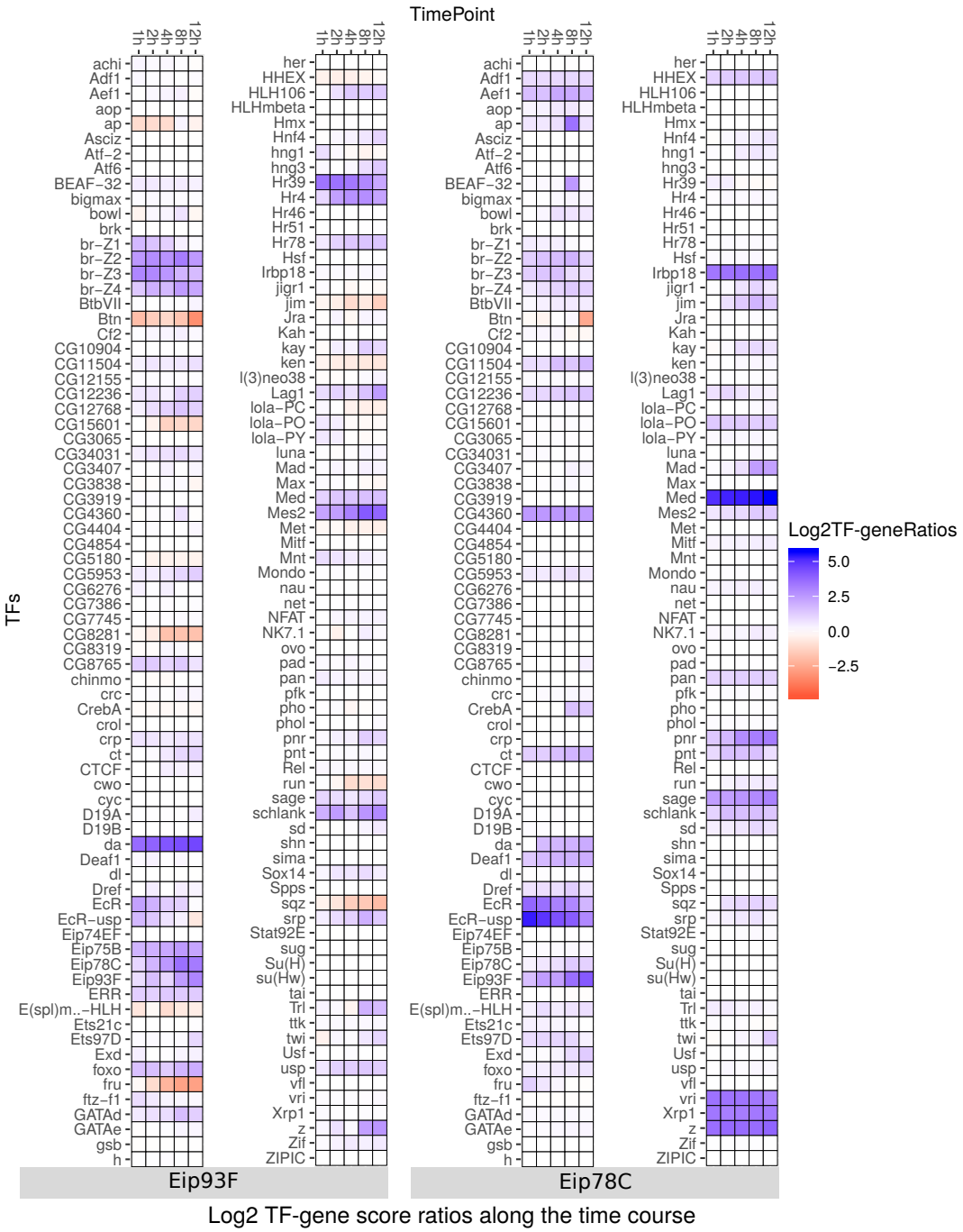

Figure 24: TF-gene score ratios for Eip93F and Eip78C.
X-axis: time points. Y-axis: TF-gene score ratios. The blue color denotes a TF-gene score ratio bigger than 1, whereas the red color denotes a TF-gene score ratio lower than 1. The darker the color, the bigger the difference between TF-gene scores of UTC and TF-gene scores of the depicted time point.

TF-gene score ratios. A classifier can be done by applying a threshold on the probability of the outcome.

We obtained the following average performances along the time course:

- accuracy: 0.677
- F1-score on upregulated genes: 0.693
- F1-score on downregulated genes: 0.659

The accuracy is defined as

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

where TP are true positives, TN true negatives, FP false positives and FN false negatives. The F1-score is defined as

$$F_1 = \frac{2TP}{2TP + FP + FN} \tag{3.2}$$

or equivalently as the harmonic mean of precision and recall.

The mathematical interpretation of the logistic regression coefficients is the following. If a TF has a positive estimated coefficient, it means that increasing the ratio of TF-gene score increases the probability of having upregulation. Viceversa if a TF has a negative estimated coefficient, it means that increasing the ratio of TF-gene score increases the probability of having downregulation. However, since logistic regression models probabilities and not $log_2(FC)$ of genes, we stick to the conservative interpretation [Durek et al., 2016] that coefficients different from zero represent TFs that are important for differential expression in the ecdysone response, and the absolute values of coefficients represent the importance that TFs have.

What is the difference between the information given by the logistic regression and the information given by the linear regression? The linear regression is fitted on *static* data, whereas the logistic regression is fitted on *differential* data. In fact, the independent variables of the linear regression are the TF-gene scores *in each* time point, and the dependent variables are the gene expression values measured *in each* time point. This means that the estimated linear regression coefficients suggest the functionality of TFs in the time points, but they do not tell us whether TFs are responsible for the observed differential regulation, since each fit involves only a specific time point. By contrast, the independent variables of the logistic regression are the TF-gene score ratios *between* each time point and UTC, and the dependent variables are the binarized $log_2(FC)$ of gene expression measured *between* each time point and UTC. This means that the estimated logistic regression coefficients suggest whether each TF is responsible for the observed differential regulation, and this is possible because each fit involves two time points.

Figure 25 shows the estimated regression coefficients. Overall, we can see only a few TFs predicted as important at 1h. All of them are still predicted as important after 2h, with the addition of several other TFs. After 4h, the set of involved TFs changes, with

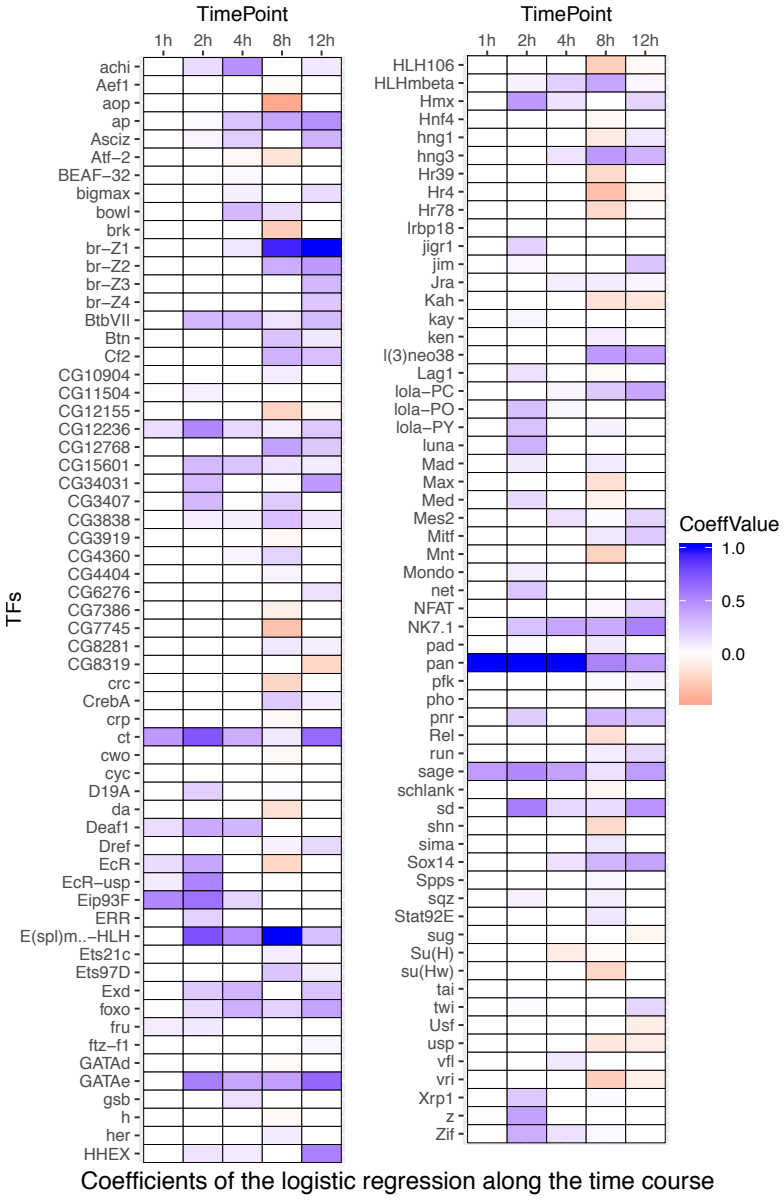

Figure 25: Estimated logistic regression coefficients along the time course for each TF.

X-axis: time points. Y-axis: TFs with at least one estimated coefficient different from 0. The blue color denotes a positive estimated coefficient, whereas the red color denotes a negative estimated coefficient. The darker the color, the stronger it is suggested that a TF is responsible for the observed gene regulation.

new players replacing the previous ones. As expected, EcR and Eip93F are suggested as important in the early time points. Br is predicted to have an effect after 4h, with the importance constantly increasing across all isoforms. Pan is known to be an activator in wing development [Schweizer et al., 2003], but not to be involved in the ecdysone response. Other TFs correctly predicted to be involved in ecdysone are Sox14 [Ritter and Beckstead, 2010] and foxo [Koyama et al., 2014]. Interestingly pnr and hng3, the most enriched TFs in the motif enrichment analysis (figure 14), are predicted to be important in the response. They are not known to be involved in the ecdysone response and therefore interesting TFs to further investigate.

3.3.5 Localization of active TFBSs via digital genomic footprinting could have given more precise TF-gene scores

Originally, digital genomic footprinting was intended to be employed in the work reported in this thesis for the identification of all the bound TFBSs. However, for all the reasons mentioned in 1.2.4, the work was carried out at a regulatory region resolution, and PWM-based predictions were employed to detect TFBSs in accessible regulatory regions. At the time of writing this thesis, neither the enzymatic cleavage bias of DNase I nor the bias of Tn5 are known to affect the identification of regulatory regions. Moreover, genome-wide correlation and correlation in detected peaks between DNase-seq and ATAC-seq are very high (figure A.9), therefore the results presented in this document are assumed to be not affected by the bias that manifests at single bp resolution.

Before we decided to complete the project described in this thesis at a regulatory region resolution, we made several efforts in the attempt to mitigate the issues of digital genomic footprinting. In particular, we decided to pursue improvements of the DNase-seq protocol. We tested four different modifications. The first one is formaldehyde fixation, to block the proteins on the DNA prior to digestion. The second one is the usage of permeabilized cells instead of nuclei extraction, to keep chromatin as unaltered as possible. The third one is to limit the digestion step in the DNase-seq protocol, to try to capture the TFs that have a short residence time. The fourth one is a more stringent size selection step, to capture only very short fragments. In our preliminary experiments permeabilized cells and stringent size selection have been tested in a single assay.

To test the effectiveness of these modifications, we have established an analyses pipeline. The main idea is to qualitatively and quantitatively assess the impact of these modifications on the distribution of cuts sites at TFBSs that lie within open regions of as many TFs as possible. For each TF we generate a heatmap that shows the distribution of cut sites for every predicted TFBS, based on available PWMs. Moreover, for each TF we compute the mean of the distributions of cut sites of all the TFBS and we generate a plot that qualitatively shows the shape of enzymatic activity. We call this qualitative shape ‘cut signature‘ of a TF. To quantify the impact of these modifications, we have developed a score that represents the depth of a footprint at each TFBS. We have conducted preliminary analyses on a set of 18 TFs. Results for the TF CTCF are shown in figure 26. The

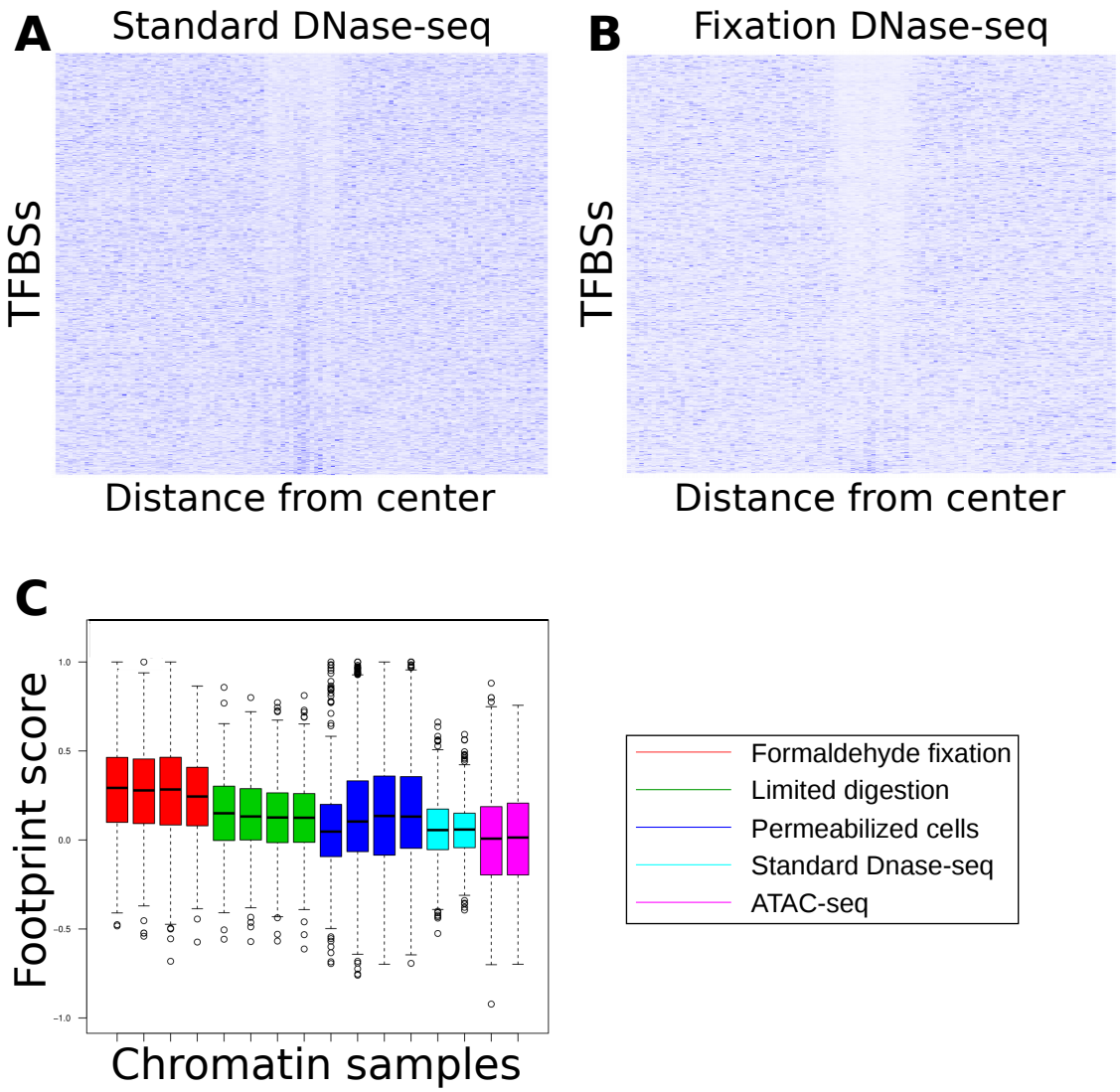

Figure 26: Footprint score distribution of different DNase-seq protocol modifications for the TF CTCF.
(A) DNase-seq activity at CTCF TFBSs with the standard protocol. Each row of the heatmap represents a single TFBS. The darker the color, the more cut sites. Rows are sorted by footprint score. (B) DNase-seq activity at CTCF TFBSs with the fixated protocol. Each row of the heatmap represents a single TFBS. The darker the color, the more cut sites. Rows are sorted by footprint score. (C) Distribution of footprint scores for all the accessibility assays tested. Red: DNase-seq with formaldehyde fixation. Green: DNase-seq with limited digestion. Blue: DNase-seq with permeabilized cells Cyan: DNase-seq standard protocol. Magenta: ATAC-seq standard protocol.

preliminary results from this set suggest that formaldehyde fixation helps in the retrieval of footprints, but further validation is needed.

In a separate investigation, we examined the cut bias of the DNase I enzyme to improve footprint detection in DNase-seq data. It has been reported in the literature that using a hexamer model to correct the cut bias is the best compromise between accuracy and speed [Gusmao et al., 2016, Sung et al., 2016, Vierstra and Stamatoyannopoulos, 2016]. However, it is possible to compute the hexamer model in two ways. The first way is to use a naked DNA control sample. In a naked DNA control sample there are no proteins that block the DNase I to cut wherever it wants. In this way the cut sites represent the DNase I cutting bias. The second way is to use the open chromatin regions. The advantage of this way to compute the hexamer model for the bias is that there is no need for a separate experiment. The cut sites here represent both DNase I cutting bias and constraints given by the environment, for example steric hindrance. To investigate the cut bias, we produced naked DNA controls using three different protocols. We calculated the hexamer bias model using all three and using open chromatin regions. We then compared the differences. The results show that DNase I bias is not strand specific and it is not affected by size selection steps in the DNase-seq protocol. However, the bias is dependent on the protocol used to produce the naked DNA sample, even though the most enriched and least enriched hexamers seem stable. By contrast, hexamers bias models calculated on open chromatin regions are similar, even among different time points in the ecdysone response cascade and among different tissues. The bias model are different also between naked DNA sample and open chromatin regions. There is no clear understanding of this phenomenon and we think that improvements on the experimental protocol will provide us with novel insights.

Ultimately, if we would have been able to use digital genomic footprinting to define the features used in the linear models, we could have obtained more precise results, since the TF-gene scores would have been computed using experimentally mapped TFBSs, as opposed to predicted TFBSs within regulatory regions. Moreover, it is of general interest to precisely detect binding events with a single assay, be it DNase-seq or ATAC-seq. For these reasons, it would be interesting to further pursue our preliminary work on digital genomic footprinting.

Chapter 4

Discussion

4.1 Expression and accessibility dynamics in ecdysone response in S2 cells

In this work, we studied the response to the hormone ecdysone during Drosophila development. To do so, we used the in-vitro paradigm of S2 cells stimulation. The system was shown to be responsive by Katja Frühauf. In her PhD thesis, she monitored ecdysone-stimulated S2 cells and reported dramatic morphological changes, with cells losing their round shape, growing in size and growing structures similar to filopodia, while starting to differentiate.

By using her nascent mRNA RNA-seq data, which allow a better quantification of the differential synthesis of mRNA during ecdysone stimulation, and integrating it with DNase-seq data, we analyzed the relationship between the regulation of gene expression and chromatin changes upon ecdysone stimulation. Our data have an unprecedented time resolution, with 6 time points captured in a 12 hours span, enabling us to capture the very early response of S2 cells to ecdysone and to have detailed profile of the dynamics of the response.

When we correlated the $log_2(FC)$ of differential genes expression and differential peaks accessibility across different time points, we noticed a great similarity between the behavior of the chromatin and the behavior of expression. This suggests a mechanism in the response where the direction of gene regulation and the direction of chromatin changes (e.g. chromatin opening and upregulation - chromatin closing and downregulation) are generally linked and concordant. Moreover, there is a noticeable imbalance towards upregulation/opening in very early time points, which fades away in late time points. This suggests a very fast reaction to the ecdysone stimulation, with cells immediately opening regulatory regions and upregulating genes needed to deal with the new environmental conditions.

To analyze this similarity more deeply, we assigned a target gene to each differential peak. As we mentioned in 3.1.3 this is not a trivial task, nevertheless we think that the assignment based on the minimization of the distances between differential peaks and

TSSs is a good approximation of the real connections between genes and regulatory regions. Further strategies of assignment will be discussed in 4.3.

We correlated the $log_2(FC)$ of differential peaks and their assigned target genes. Again, their behavior is generally concordant, with opening regulatory regions favoring upregulation of their regulated genes and closing regulatory regions favoring downregulation of their regulated genes. This behavior is even more pronounced when we restricted the analysis to differential promoters, whereas restricting the analysis to differential enhancers did not alter the results. These data suggest a mechanism where induced inaccessibility of regulatory regions leads to repression of target genes, whereas induced accessibility leads to activation. However, a considerable part of closing regulatory regions are associated with upregulated genes. This indicates a more complicated role for the repression of chromatin, that could be fine tuning the amount of synthetized mRNA, or it could be multiple layers of regulation for some genes in the ecdysone response.

Given this linear relationship between regulation and chromatin openness of regulatory regions, we analyzed whether the number of associated differential peaks played a role in the expression of target genes. Indeed, genes with more associated opening peaks are more strongly activated, even when removing promoters from the analysis, which could bias the results given their more linear relationship. By contrast, expression of genes with associated closing peaks is repressed, but the number of associated closing peaks does not play a role in the amount of repression. This suggests some synergistic interaction between activating enhancers of the same target genes, whereas repressing enhancers do not cooperate in shutting down gene expression. For instance, an additional opening enhancers could recruit a co-activator that further increases gene expression, whereas closing enhancers simply hinder binding.

Given the temporal resolution that we achieved in our experiments, we analyzed the dynamics of the differential genes and differential peaks. Using hierarchical clustering with cosine distance to measure the differences between the profiles, without taking into account their absolute magnitude, we defined 3 clusters for the profiles of differential peaks and 4 classes for the profiles of differential genes. The 3 clusters for the differential peaks represent closing peaks, opening peaks and early opening peaks. Analogously, these clusters could be defined also for the differential genes, with an additional cluster representing early downregulated - late upregulated genes. Therefore, S2 cells are responding to the ecdysone stimulus with an immediate upregulation of some genes, followed by upregulation and downregulation of other genes. Accessibility dynamics follow a similar behavior. Together, this rules out the possibility of a periodical kind of response, where genes are continuosly increasing and decreasing their expression in a sinusoidal fashion, and confirms that S2 cells are responding to the stimulus with a typical response to an environmental perturbation.

Since S2 cells respond to ecdysone stimulation with a response in line with an environmental perturbation, we could use ImpulseDE2 [Fischer et al., 2017] to have more faithful models of expression dynamics and accessibility dynamics. ImpulseDE2 allows better detection of differential genes and differential peaks, because it analyzes and models all the time points at once. Moreover, it is able to classify the dynamics in one of the four classes Tn-U, Tn-D, Tt-U and Tt-D. A GO terms enrichment analysis on the genes classified as

Tn-U or Tn-D recovers terms that were found by Katja Frühauf, confirming the quality of the classification. In the Tt-U and Tt-D classes, all the dynamics are switching direction at 4h, in what it seems to be a key turning point in the ecdysone response. Moreover, the proportion between different classes between expression and accessibility are very similar.

Given this similarity, we employed the Jaccard index to measure how much differential genes and target genes of differential peaks classes overlap. As expected, differential genes and target genes of differential peaks belonging to the same class agree between each other. This gives strength to our hypothesis of a general linear mechanism between opening or closing chromatin and upregulation or downregulation of expression. However, genes classified as Tn-U agree with differential peaks classified as Tn-D and Tt-U, suggesting a more refined regulation, possibily with more than one layer of interactions, where closing chromatin is not associated with repression but has a more sophisticated role. This phenomenon is even more pronounced when we considered only TFs, with the agreement being higher between Tn-U TFs and Tt-U peaks than between Tn-U TFs and Tn-U peaks. This suggests that TFs that are involved in the ecdysone response have even a more complex regulation, in line with the behavioral and morphological changes that ecdysone triggers.

4.2 Motif enrichments suggest TFs thesauri

After a thorough analysis of the dynamics and the relationships between chromatin and regulation after ecdysone stimulation, we selected TFs that respond to the stimulus and conducted a motif enrichment analysis to determine which ones could be key players. 3 isoforms of br are enriched in several classes of differential peaks. In particular, br-Z1 is enriched in all classes and br-Z2 is highly enriched in Tn-D peaks, highlighting once more that br is a fundamental TF in the ecdysone response in S2 cells. Srp has been suggested to have a role in ecdysone-induced enhancers in S2 cells [Shlyueva et al., 2014]. In line with this, it is highly enriched in the Tn-U class. However, srp is enriched in all classes, suggesting a more general role of this TF. Foxo, another highly enriched TF in our analysis, is involved in ecdysone biosynthesis [Koyama et al., 2014]. Among the enriched TFs, hng3, pnr and CG5953 have not been previously associated with ecdysone. We will discuss foxo, hng3 and pnr in more detail in section 4.3, since they are also suggested as being responsible for regulation from the logistic regression model. Overall, we were able to suggest TFs that could have a key role in the ecdysone response in S2 cells.

Analogously, we selected TFs that potentially respond to the stimulus in larval tissues and conducted tissue-specific motif enrichment analyses. After filtering the results to have TFs that appears at most once across all the tissues, we were able to suggest TFs that could be involved in determining the fate of each tissue after ecdysone stimulation.

4.3 Linear models deepen understanding of S2 cells ecdysone response

We decided to get further insights on the ecdysone stimulation by using linear models to predict expression levels. In particular, we used linear regression to suggest functionalities of TFs, and logistic regression to suggest TFs responsible for regulation in the response. Both models use elastic net regularization to ease the interpretability of the results, by making them sparse. A limitation of linear modeling is already in its name. Reality is never linear, and by using a linear model we already know that it is an approximation. Nevertheless, linear models are very useful, because they provide an easy interpretation of the results. If we were to use a non-linear model, we may have obtained a model that is more adherent to reality, but we may not have been able to interpret results as easily as we do with linear models.

Another limitation of our linear approach is to assume that interactions between TFs are purely additive. In other words, we are excluding from our modeling that TFs that cooperate synergistically may increase expression much more than simply adding the single effects of each TF. In a linear setting, one could think to model this by defining an independent variable for each possible interaction between TFs. However, this leads to the explosion of the number of variables. For example, in our case, if we wanted to model all the interactions between pairs of TFs, we would have needed to add $148^2 = 21904$ independent variables, making impossible the fitting of the model.

TF-gene scores represent the *affinities* that TFs have for each gene, therefore their definition is a fundamental part of the modeling effort. TF-gene scores depend on the association between peaks and target genes. For this reason, we tested 2 further strategies in addition to the assignment by nearest TSS. One strategy assigns a *region of influence* to each gene, and peaks overlapping it are assigned to the relative gene. Each *region of influence* is at least as small as the gene itself, but they are usually bigger. The other strategy, which has been used in other works [Ouyang et al., 2009, McLeay et al., 2012, Schmidt et al., 2016], assigns a window centered on the TSS to each gene, and peaks overlapping it are assigned to the relative gene. The size of the window is defined by the user.

The assignment by windows centered on TSSs gave the best performances when we used it to define the TF-gene scores to fit linear regression. While the assignment by nearest TSS is a good approximation when used to characterize the relationships between differential genes and differential peaks, it is not well suited when modeling expression levels. It could be that the assignment is too restrictive, because each peak gets associated to exactly one gene, not allowing for uncertainty in the effect of enhancers that have a similar distance to TSSs of different genes. Another disadvantage is that peaks that are far away from any TSS get nevertheless assigned to a gene. This could penalize the accuracy of the TF-gene scores, even though it is mitigated by the distance exponential downweighting. It is more difficult to speculate on the reasons of the lower performances of the *region of influence* approach. Exactly as the window centered on TSS approach, each peak can get assigned

to several genes in dense genomic regions, and peaks that are far away from any gene do not get assigned. However, in intronic regions with intronic genes, the 2 strategies behave differently. *Region of influences* assigns peaks that are near a TSS of an intronic gene also to the gene whose intron belongs to, whereas the window on TSS strategy does it only when the TSS of an intronic gene is sufficiently close to the TSS of the gene whose intron belongs to. Even though intronic regulatory regions usually regulate the gene that contains them [Kvon et al., 2014], it could be that if they lie near a TSS of an intronic gene they regulate exclusively that gene. This would cause the *region of influences* strategy to make erroneous assignments in this case, consequently worsening the TF-gene scores.

Factoring in the TF-gene scores the expression levels of TFs improved the performances with respect to a TF-gene scores definition that does not use them. If a gene has regulatory region near its TSS with strong affinity for a particular TF, the TF-gene score for that gene will be high. However, if the TF is not expressed in the system under study, it can not have an effect, therefore factoring the expression levels make the TF-gene scores closer to reality. Conversely, if a TF-gene score is very low, but the TF is highly expressed in the system, such TF will have an impact on the regulation of the gene, and factoring the expression level of the TF will increase the TF-gene score to capture this phenomenon. Interestingly, factoring the mean accessibility of peaks did not give the best performances. It could be that after a peak is 'open enough' to allow for binding, the level of accessibility does not play a role in the regulation of transcription, and making the TF-gene score proportional to it is detrimental. However, by looking at a greater set of top results, it is not clear if, in general, factoring the mean accessibility helps the fit or if it does not. This is a point that it is worth further investigations.

The estimated coefficients of the linear regression generally recover known functionalities of TFs. EcR is predicted as repressor without ecdysone and as activator after ecdysone stimulation. Eip75B is predicted as repressor. Sage is predicted as activator. Other correctly predicted TFs are sd, which acts as coactivator when it forms a complex with Yki [Goulev et al., 2008, Wu et al., 2008, Zhang et al., 2008a, Zhao et al., 2008], and pad, which is a repressor of achaete-scute during bristle development [Gibert et al., 2005].

The ratios between TF-gene scores of 2 different time points represent the variation of *affinities* that TFs have for each gene. To define them, we used the TF-gene scores definition that gave the best performances. For each time point comparison, the estimated coefficients of the logistic regression suggest which TFs could explain the observed regulation. Over time, estimated coefficients behave as we would expect, with a few TFs considered important in early time points, whereas after 4h their number increase considerably. EcR and Eip93F are suggested as important in early time points, whereas br has an effect after 4h. Pan, predicted as very important, is not known to be involved in the ecdysone response and it acts as activator in wing development [Schweizer et al., 2003], in agreement with the prediction from the linear regression. Other TFs are correctly predicted as involved in the ecdysone response, such as Sox14 [Ritter and Beckstead, 2010] and vri [Gauhar et al., 2009].

Interestingly, foxo, hng3 and pnr are predicted as important. These 3 TFs were already highlighted in the motif enrichment analysis, where they are all highly enriched in both

Tn-U and Tn-D peaks. Foxo is involved in ecdysone biosynthesis, but with our results we speculate that it could have a more central role. Pnr is involved in the development of imaginal discs and nervous system, but has never been associated with ecdysone. Little is known about hng3, and in particular it is not known to be involved in ecdysone responses. For these reasons, we suggest foxo, pnr and hng3 as candidate TFs for further investigation of the ecdysone response in S2 cells.

4.4 Conclusions and outlook

In this thesis we used the ecdysone response as paradigm to study regulation of transcription. By doing so, we have also deepened the understanding of the ecdysone response itself, and we have seen to which extent and resolution DNase-seq data can be used to model expression.

While we have obtained good correlations between predicted expression and measured expression in the regularized regression model, clearly there is room for improvement. A limitation comes from the linearity of the model, because regulatory events do not influence expression in a linear fashion. Another limitation could come from the associations between regulatory regions and target genes. We have tested some heuristic assignments and used the window on TSS approach, which gave the best results. However, having genome-wide annotations that identify enhancers and link them with the regulated gene would greatly improve predictions in any model. Moreover, one could further improve predictions if every protein-protein interaction were annotated, by including interaction terms in models. Unfortunately, these are not trivial task, and they will not be completed in the near future.

In our linear modeling we have modeled each time point independently. A possible improvement could be to model all the time points jointly, and exploit the fact that they are not independent. Moreover, additional downstream analyses on the TF-gene score ratios are possible. For example, one could cluster them and see whether sets of target genes show common regulators, or if it is possible to see general common regulatory patterns and suggest co-regulating TFs.

So far, we have applied the linear models only to the ecdysone response in S2 cells, due to the lack of expression data in the larval paradigm. It would be interesting to apply the same models in the more complicated in-vivo paradigm, to compare the results and gather additional insights on the ecdysone response in larvae.

For his PhD project Andrea Ennio Storti perturbed the response to ecdysone in S2 cells by knocking-down EcR and br, and revealed how the regulatory and the expression responses changed. A possibility to use these data would be to check whether an in-silico knock-down in the linear models give similar results as the experimental knock-down. Further interesting knock-downs could be done, for example on the TFs foxo, pnr and hng3, because their motifs are highly enriched in differential peaks in S2 cells and they are predicted as important to explain differential expression by the logistic regression model. Knock-down of at least one of these TFs could show a tremendous effect on the system, opening the way to add it to the TFs that are key players in the ecdysone response.

Chapter 5

Further contributions

5.1 Regional differences in enhancer accessibility in Drosophila blastoderm

This work has been submitted for publication as:

ATAC-seq reveals regional differences in enhancer accessibility during the establishment of spatial coordinates in Drosophila blastoderm

Marta Bozek, Roberto Cortini, Andrea Ennio Storti, Ulrich Unnerstall, Ulrike Gaul, Nicolas Gompel

Contribution: I mapped, processed and called peaks on all the ATAC-seq samples generated in this study.

5.1.1 Introduction

Gradients of concentration of several activating and repressing TFs along the axis of Drosophila embryos are required for a proper spatial activation of enhancers, which regulate target genes responsible for the correct patterning. However, the relation between chromatin accessibility and enhancer activity is not well understood, and it is not known whether accessibility is uniform across the embryo or it varies along the axis.

5.1.2 Results

Measurement of accessibility along embryonic axis with ATAC-seq shows that one quarter of the accessible genome has significant regional variation. Moreover, accessibility changes correlate with the regulatory activity of enhancers (figure 27). In regions of the embryo where an enhancer receives activating TFs and promotes transcription, its accessibility is higher with respect to regions where it receives repressing TFs.

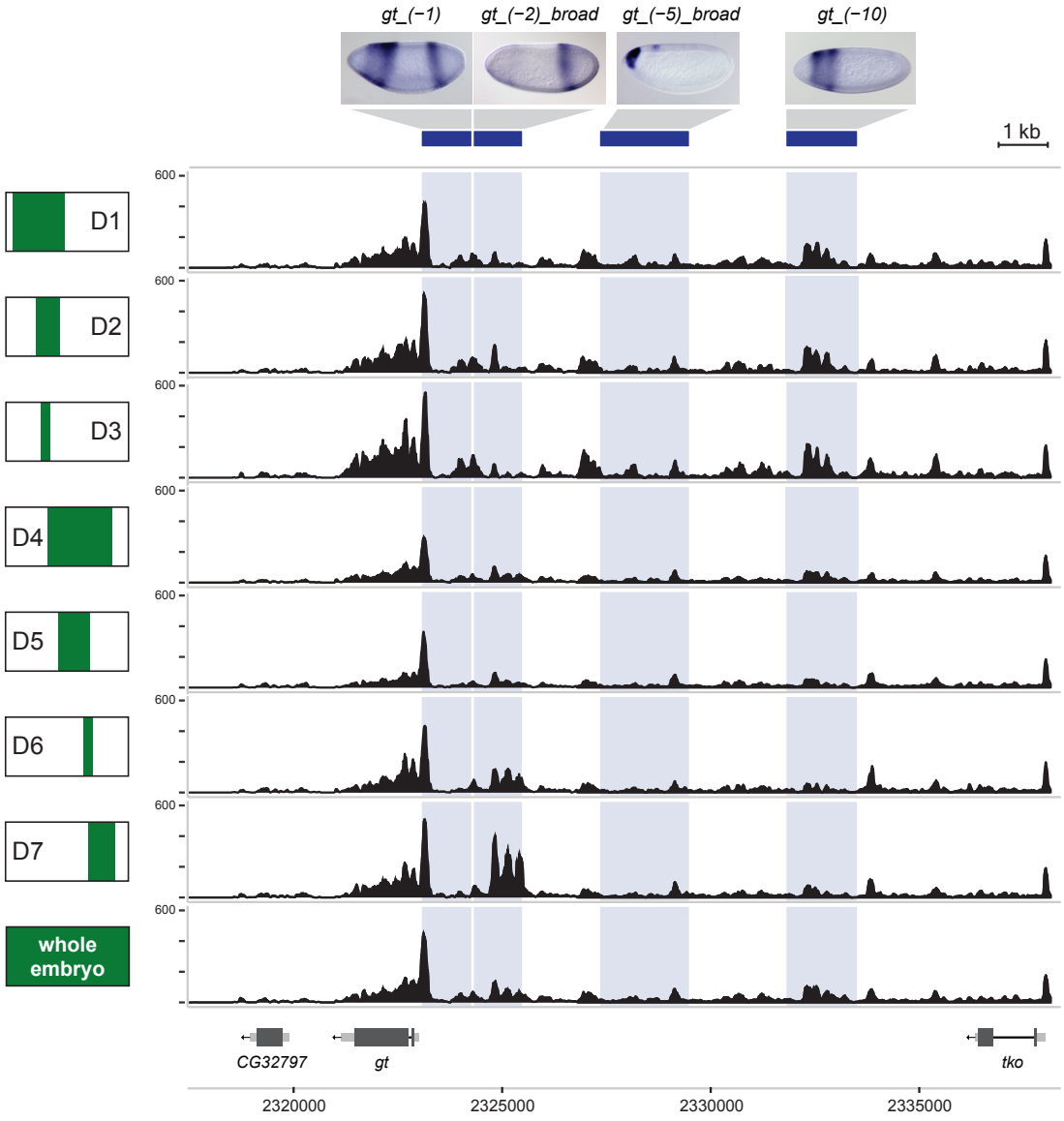

Figure 27: Regional differences in chromatin accessibility illustrated at the giant locus.
Accessibility profiles of individual tagged domains and a whole-embryo control at the locus of giant, a gene of the anter-posterior patterning network of the gap class. Tracks show normalized coverage of 1-100 bp ATAC-seq fragments, smoothed over a sliding window of 15 bp. Antero-posterior positions of the profiled domains are indicated schematically on the left (green shading). Blue bars and underlying shaded regions indicate coordinates of known giant enhancers. Spatial activity of each enhancer in blastoderm embryos is illustrated above. Horizontal axis represents genomic coordinates along chromosome X.

5.1.3 Conclusions

Chromatin context plays a role in the spatial regulation of enhancers responsible for the patterning of the axis. Differential accessibility is a signature of differential regulatory activity and can potentially serve as a metric for de novo identification of enhancers patterning complex tissues.

Appendix A

Supplementary figures

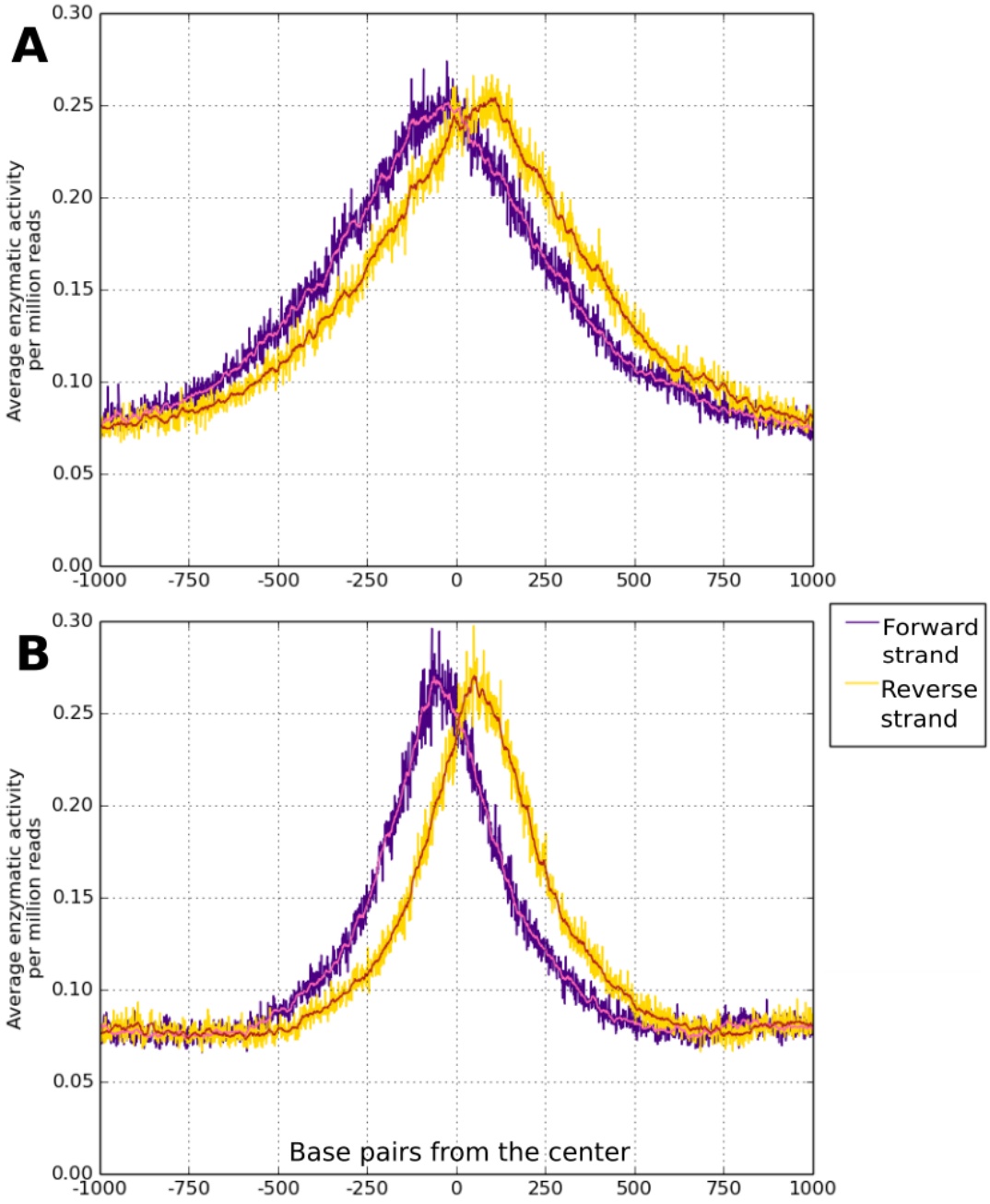

Figure A.1: Average cut frequency of our DNase-seq data in known accessible regulatory regions.
Regulatory regions found in [Arnold et al., 2013] using (A) DNase-seq and (B) STARR-seq data were used to compute the average cut frequency of our data, which show enrichment in regions known to be accessible. Purple/pink: cut sites aligned to the forward strand. Yellow/red: cut sites aligned to the reverse strand.

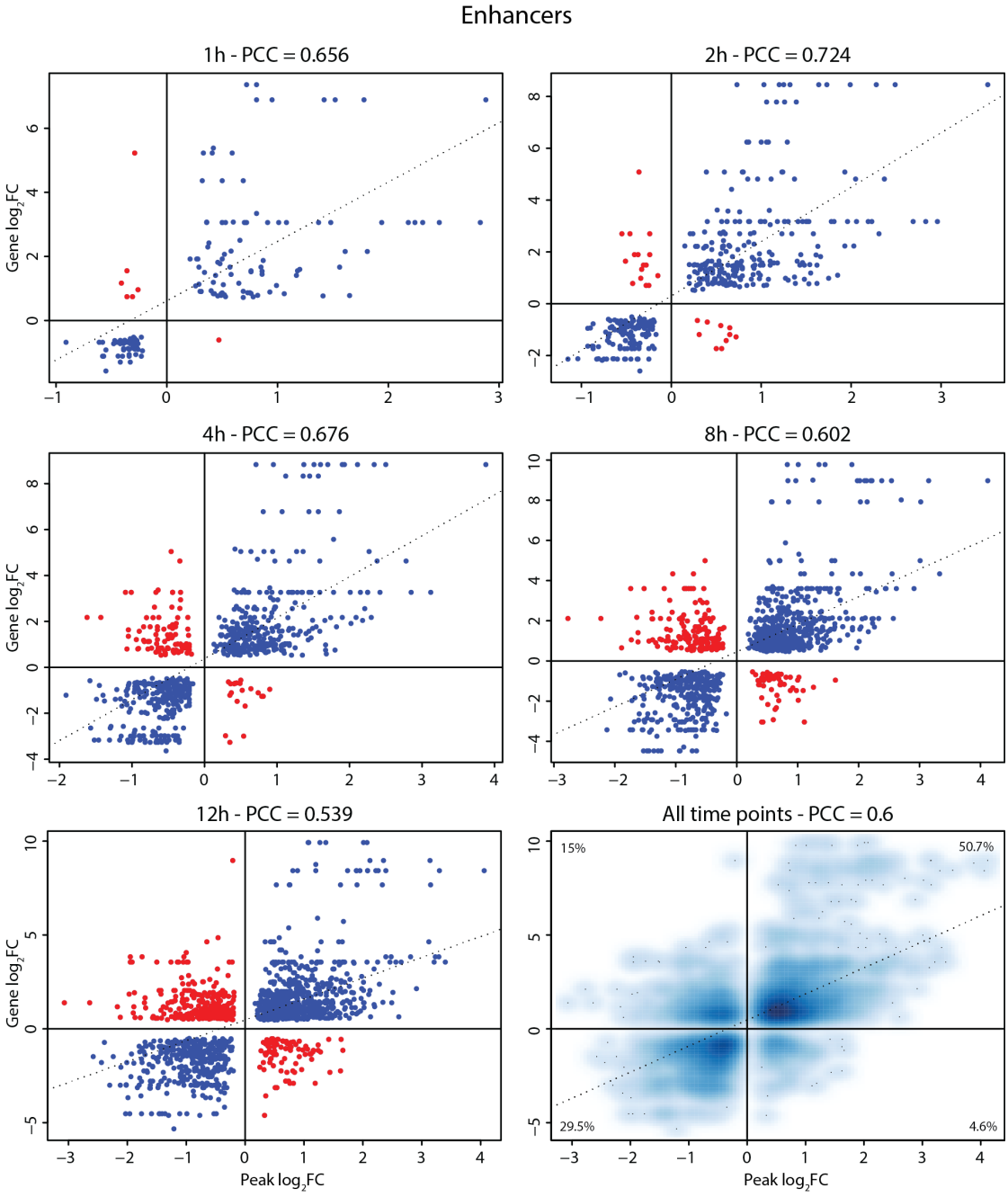

Figure A.2: Scatterplots between $log_2(FC)$ of differential enhancers and $log_2(FC)$ of their target genes.
For each time point, $log_2(FC)$ of differential enhancers (x-axis) and $log_2(FC)$ of their target genes (y-axis) was correlated. Correlation values are shown above each plot. Blue dots: $log_2(FC)$ that agree in sign. Red dots: $log_2(FC)$ that do not agree in sign.

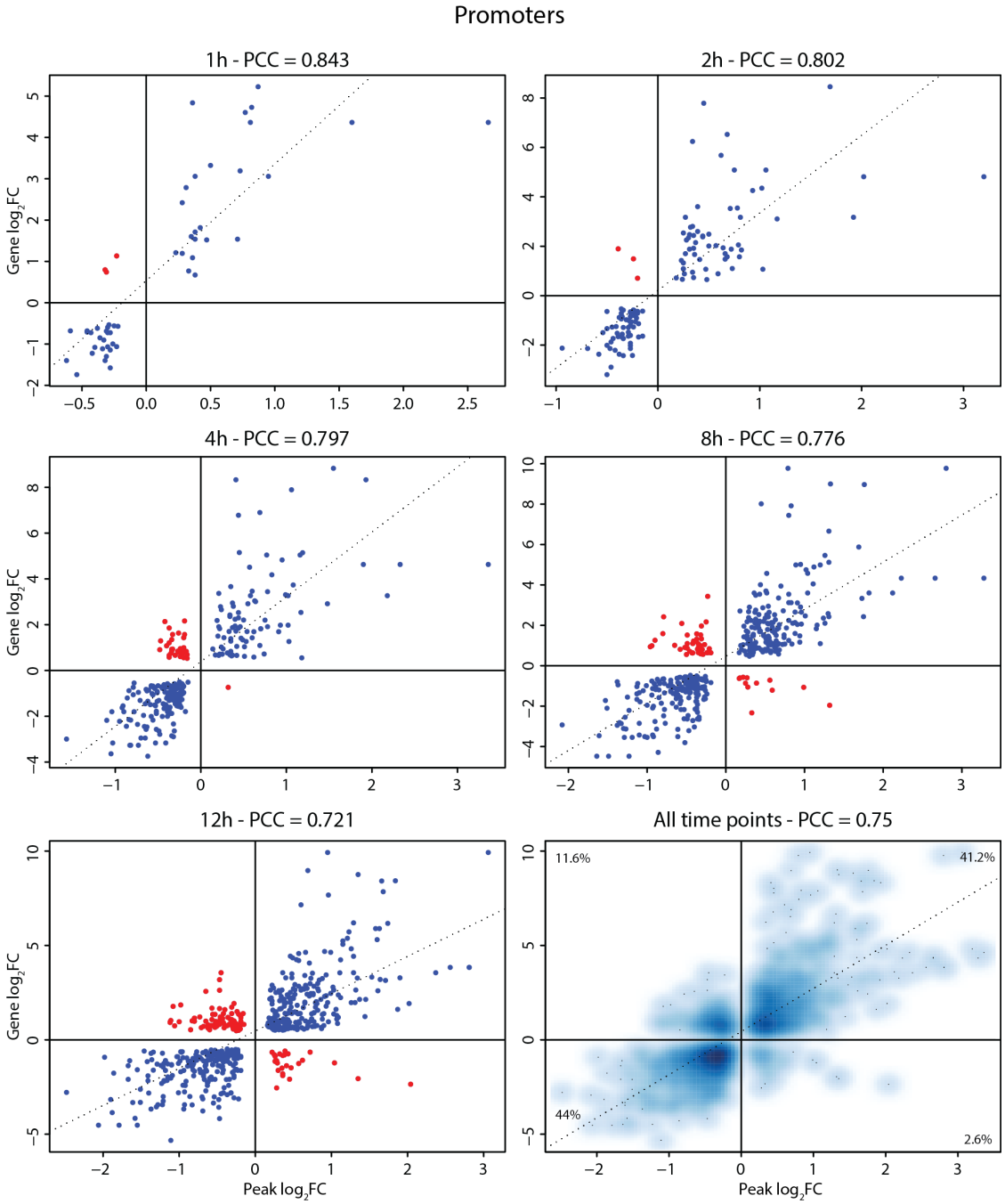

Figure A.3: Scatterplots between $log_2(FC)$ of differential promoters and $log_2(FC)$ of their target genes.
For each time point, $log_2(FC)$ of differential promoters (x-axis) and $log_2(FC)$ of their target genes (y-axis) was correlated. Correlation values are shown above each plot. Blue dots: $log_2(FC)$ that agree in sign. Red dots: $log_2(FC)$ that do not agree in sign.

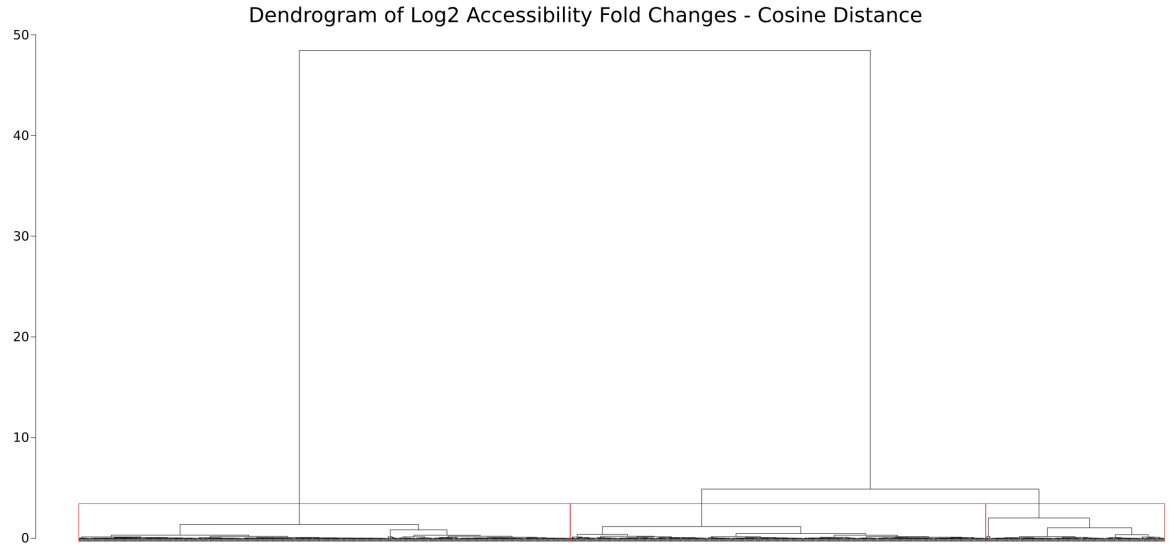

Figure A.4: Dendrogram produced by hierarchical clustering of the accessibility profiles of differential peaks over time.
Red line indicates the cut performed to obtain 3 clusters.

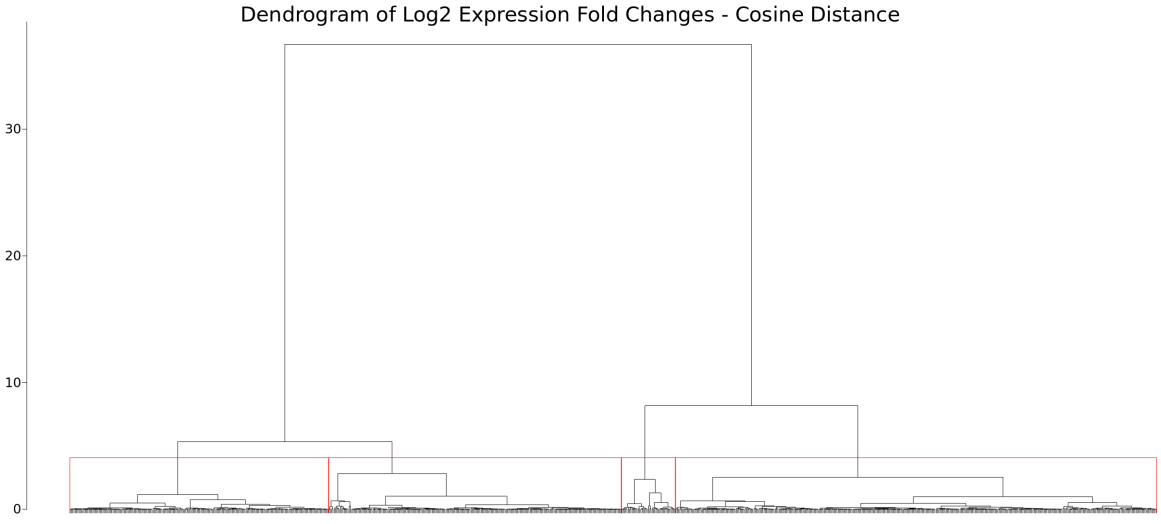

Figure A.5: Dendrogram produced by hierarchical clustering of the expression profiles of differential peaks over time.
Red line indicates the cut performed to obtain 4 clusters.

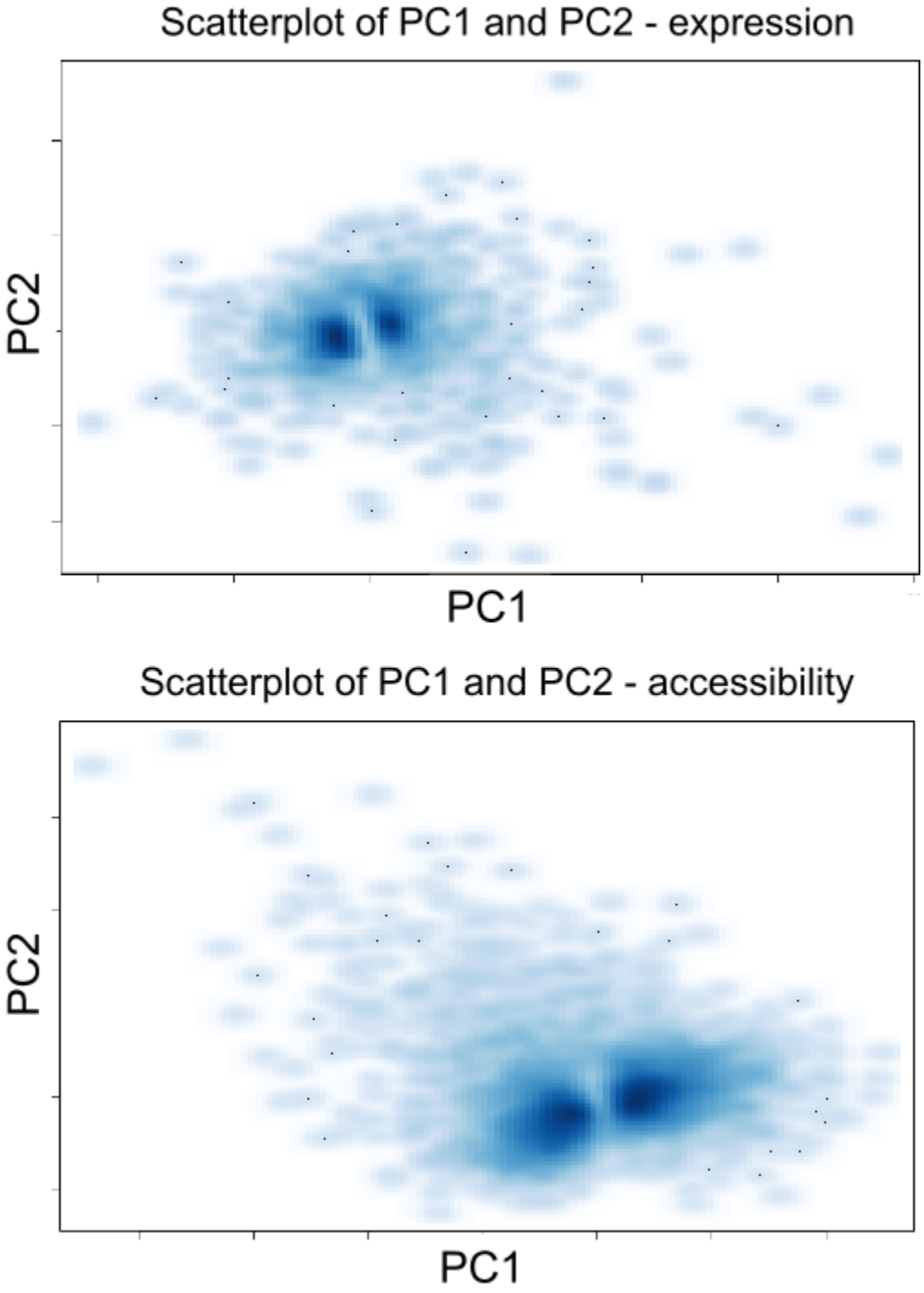

Figure A.6: PCA of expression dynamics and of accessibility dynamics. Explained variance by expression PC1: 87%. Explained variance by expression PC2: 10%. Explained variance by accessibility PC1: 86%. Explained variance by accessibility PC2: 11%.

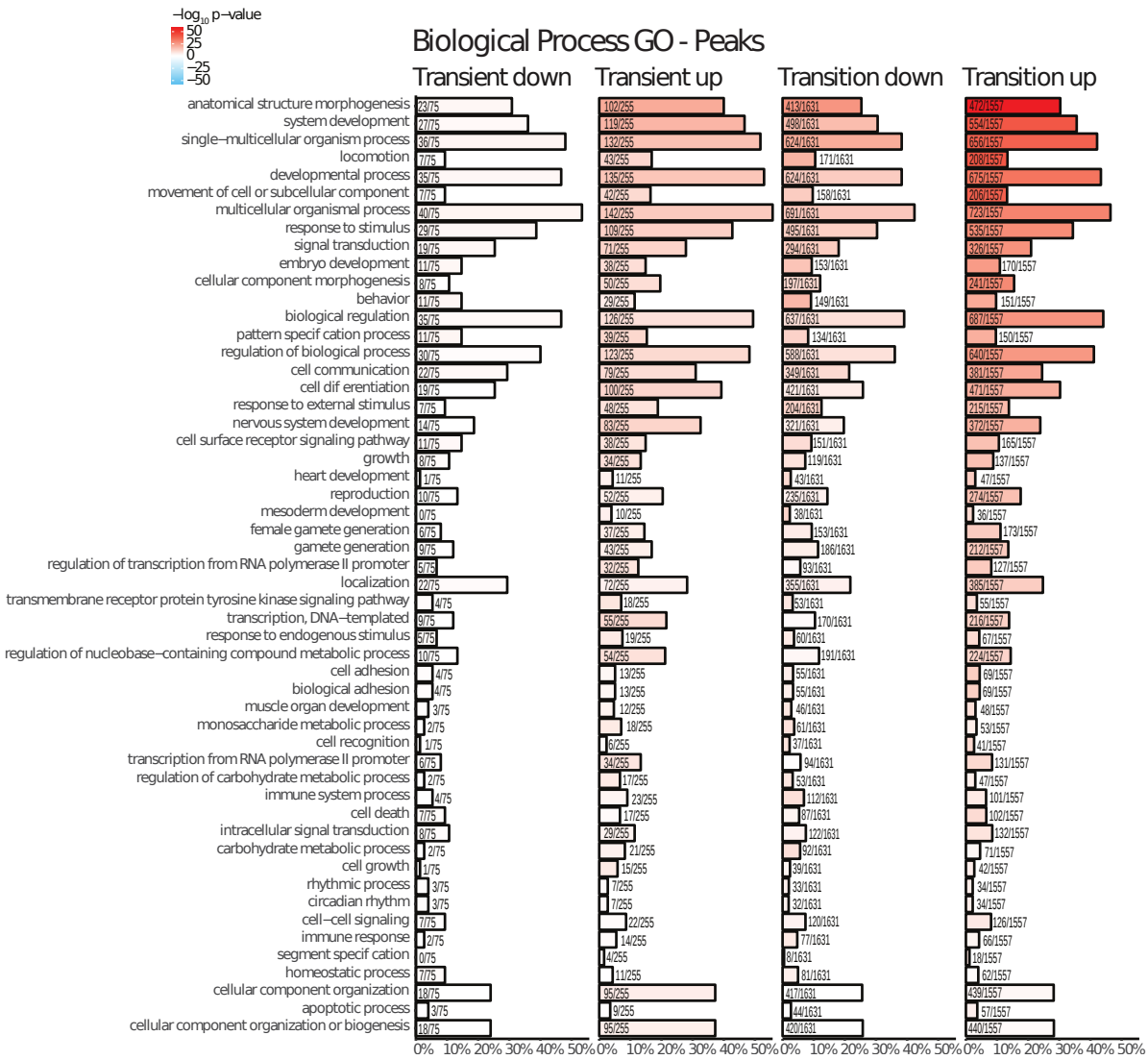

Figure A.7: GO terms enrichment analysis on target genes of differential peaks, grouped by class of differential peaks assigned from ImpulseDE2.
Bars colored in red represent enriched terms, whereas bars colored in blue represent depleted terms. Intensity of the color represents $-log_{10}(pvalues)$. Size of the bars represents the ratio (indicated on or next to the bars) between the number of genes in the class annotated to the specific term and the number of genes in the class, with the percentages scale on the x-axis.

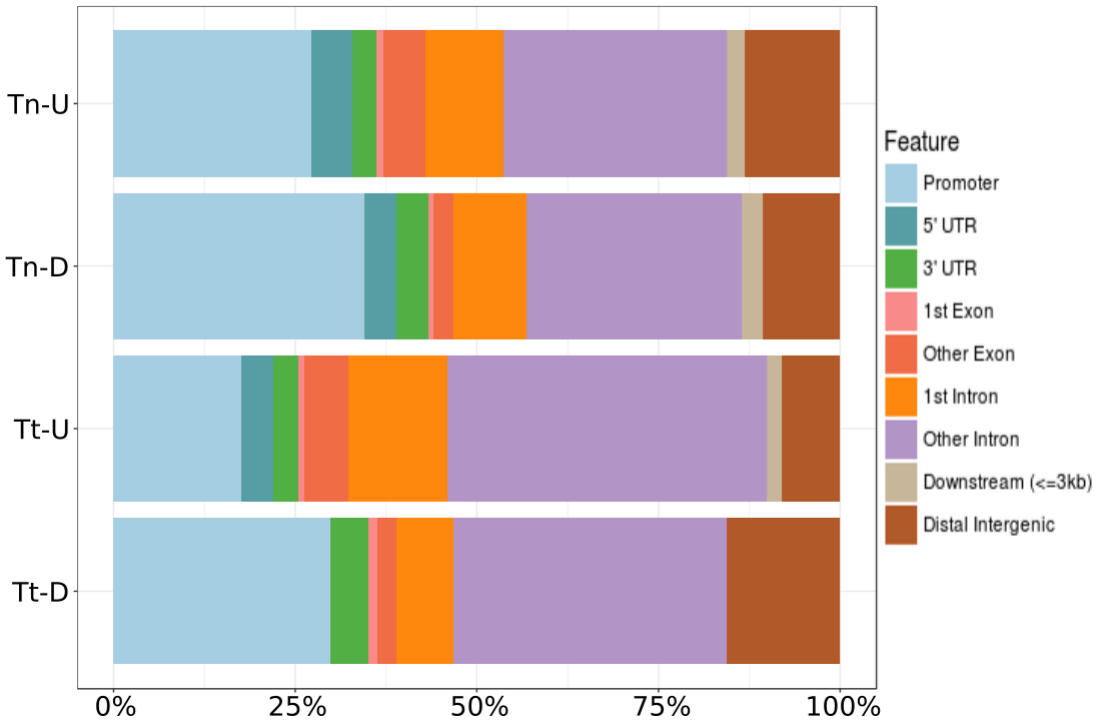

Figure A.8: Distribution of annotations of differential peaks, grouped by class. Each differential peak was annotated with the genomic feature that it overlaps. Proportions of annotations within each class are reported.

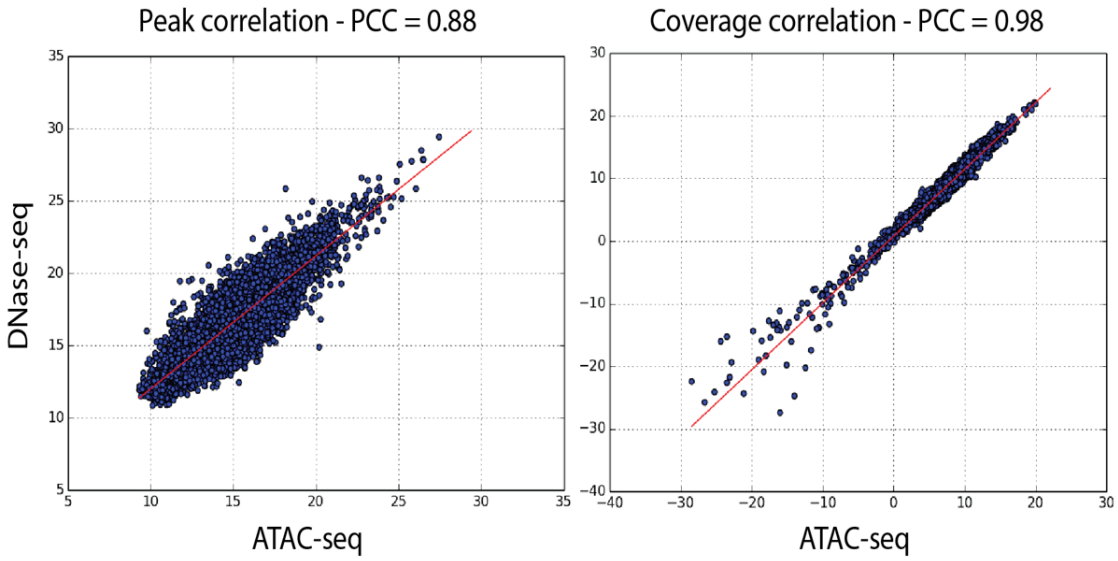

Figure A.9: Correlation between a DNase-seq sample and an ATAC-seq sample on S2 cells.
X-axis: mean ATAC-seq coverage (log scale). Y-axis: mean DNase-seq coverage (log scale). Left plot: mean coverage computed on overlapping accessibility peaks. Right plot: mean coverage computed on 10kbp bins in a genome-wide fashion.

Appendix B

Tables

br-Z1	hng3
br-Z2	Hr39
br-Z3	Hr46
br-Z4	Hr4
Btn	Hr51
CG5953	kay_Jra
chinmo	ken
dl	lola-PC
EcR	lola-PO
EcR-USP	lola-PY
Eip74EF	Max_Mnt
Eip75B	NFAT
Eip78C	pnr
Eip93F	schlank_Lag1
Ets21c	shn-F1-2
foxo	Sox14
fru	srp
ftz-f1	tai_Clk
h	twi_da
Hnf4	vri

Table B.1: List of TFs with differential behavior used for the motif enrichment in S2 cells.

abd-A	ci	Ets98B	Max	schlank
Abd-B	crc	exex	Med	Scr
ac	crol	foxo	mirr	sd
Adf1	crp	fru	Mnt	sens
Aef1	cwo	ftz-f1	Mondo	shn
al	cyc	GATAd	net	sim
aop	D	gl	NK7.1	slp1
ap	D19B	grh	nub	sob
ara	da	gsb	oc	Sox14
Atf6	Dfd	gsb-n	odd	Sox15
ato	disco	h	opa	Sp1
Awh	disco-r	Hand	Optix	sqz
bcd	Dll	Hnf4	otp	sr
Bgb	Doc2	hng3	ovo	ss
B-H1	dpn	Hr3	pad	su(Hw)
B-H2	Dr	Hr78	pan	sv
Blimp-1	dsx	Hsf	pb	svp
bowl	dysf	hth	pdm2	tai
br	E(spl)m3-HLH	inv	peb	toy
brk	E(spl)m8-HLH	jim	pho	Trl
btd	E(spl)mbeta-HLH	Kah	phol	ttk
byn	E(spl)mdelta-HLH	kay	pnr	tup
cad	E5	ken	pnt	twi
caup	ecr_usp	kni	Poxm	unpg
CG12236	eg	knrl	Poxn	Usf
CG15812	Eip74EF	Kr	prd	vri
CG3838	Eip75B	lab	Ptx1	vvl
CG3919	Eip78C	lbe	Rel	wor
CG4360	ems	lbl	retn	Xrp1
CG5953	en	Lim1	rib	z
CG8765	esg	lola	salr	ZIPIC
chinmo	Ets21C	Mad	sc	zld

Table B.2: List of TFs with differential behavior used for the motif enrichment in WD.

abd-A	Dlip3	Hr78	retn
ac	Dll	hth	rib
Adf1	Doc2	jim	sc
al	dsx	kay	Scr
ara	dysf	klu	sd
Atf6	E(spl)m3-HLH	knrl	sens
ato	E(spl)m8-HLH	Kr	shn
Awh	E(spl)mbeta-HLH	lab	sim
Bgb	E(spl)mdelta-HLH	lbl	slp1
B-H1	eg	Lim1	sna
B-H2	Eip74EF	Lim3	Sox14
Blimp-1	Eip75B	Lmx1a	Sox15
bowl	Eip78C	lola	sr
br	ems	Mad	ss
brk	en	Max	Su(H)
byn	esg	Mnt	su(Hw)
cad	Ets21C	NK7.1	sv
caup	Ets65A	nub	svp
CG3838	eve	oc	tai
CG3919	exex	odd	tin
CG5953	fru	opa	tj
CG8765	ftz-f1	otp	toy
chinmo	grh	ovo	ttk
ci	gsb	pan	tup
crc	gsb-n	pnr	Ubx
cwo	h	pnt	unpg
D	H2.0	Poxm	vri
da	Hand	Poxn	Xrp1
Deaf1	Hnf4	prd	zld
Dfd	hng3	Rel	
disco	Hr3	repo	

Table B.3: List of TFs with differential behavior used for the motif enrichment in ED.

abd-A	chinmo	ey	lola	sens
Abd-B	ci	Fer1	luna	shn
ac	crc	Fer2	Mad	Sidpn
achi	crol	Fer3	Max	sim
Adf1	crp	fkh	mirr	Six4
Aef1	cyc	foxo	Mnt	slou
al	D	fru	Mondo	slp1
aop	D19A	ftz-f1	net	slp2
ap	D19B	GATAd	NK7.1	sna
ara	da	gl	nub	sob
Asciz	Deaf1	gsb	oc	Sox14
ato	Dfd	gsb-n	odd	Sp1
Awh	dimm	gt	Oli	sr
bab1	Dlip3	h	onecut	ss
bap	Dll	Hand	opa	Su(H)
bcd	dpn	hb	Optix	sv
Bgb	Dr	hbn	otp	svp
B-H1	dsx	Hey	ovo	tai
B-H2	E(spl)m3-HLH	HGTX	pad	tap
Blimp-1	E(spl)m8-HLH	HLH4C	pan	tin
bowl	E(spl)mbeta-HLH	Hnf4	pb	tj
br	E(spl)mdelta-HLH	hng3	PHDP	tll
brk	E(spl)mgamma-HLH	Hr3	pho	toy
bsh	EcR	Hr78	pnr	Trl
btd	eg	Hsf	pnt	ttk
cad	Eip74EF	hth	Poxn	tup
caup	Eip75B	ind	Ptx1	Ubx
CG12236	Eip78C	inv	Rel	unpg
CG12605	ems	jim	repo	Usf
CG15812	en	kay	retn	vri
CG3838	erm	ken	rib	vvl
CG3919	ERR	klu	run	wor
CG4328	esg	Kr	salr	Xrp1
CG4360	Ets21C	lab	sc	ZIPIC
CG4404	Ets65A	lbl	schlank	zld
CG5953	eve	Lim1	Scr	
CG7368	exd	Lim3	scrt	
CG8765	exex	Lmx1a	sd	

Table B.4: List of TFs with differential behavior used for the motif enrichment in CNS.

abd-A	ci	Ets97D	lola	sd
Abd-B	Clk	Ets98B	luna	sens
ac	crc	eve	Mad	shn
achi	CrebA	exd	Max	slbo
Adf1	crol	exex	Med	slp1
Aef1	crp	ey	mirr	slp2
al	cwo	fkh	Mnt	sob
aop	cyc	foxo	Mondo	Sox14
ap	D	fru	net	Sox15
ara	D19A	ftz-f1	NFAT	Sp1
Asciz	D19B	GATAd	NK7.1	Spps
Atf6	da	gl	nub	sqz
Awh	Deaf1	grh	oc	sr
bcd	Dfd	gsb	odd	ss
Bgb	disco	gsb-n	opa	Su(H)
B-H1	disco-r	h	Optix	su(Hw)
B-H2	Dlip3	hb	otp	sug
bin	Dll	Hnf4	ovo	sv
Blimp-1	Doc2	hng3	pad	svp
bowl	Dr	Hr3	pan	tai
br	dsx	Hr78	pb	tin
brk	dysf	Hsf	pdm2	toy
btd	E(spl)m3-HLH	hth	peb	Trl
cad	E(spl)m7-HLH	inv	pho	ttk
caup	E(spl)mbeta-HLH	jim	pnr	tup
CG11617	E5	Jra	pnt	Ubx
CG12236	EcR	kay	Poxm	unpg
CG15812	eg	ken	Poxn	vri
CG3407	Eip74EF	klu	prd	vvl
CG3838	Eip75B	kni	Ptx1	wor
CG3919	Eip78C	knrl	Rel	Xrp1
CG4360	ems	Kr	repo	z
CG5953	en	lab	rib	Zif
CG6272	erm	lbe	salr	zld
CG6276	ERR	lbl	sc	
CG8765	esg	Lim1	schlank	
chinmo	Ets21C	Lim3	Scr	

Table B.5: List of TFs with differential behavior used for the motif enrichment in SG.

TF	L3IL opening	L3IL closing	WPP opening	WPP closing
AbdA	-27.1	Not enriched	-36.2	Not enriched
AbdB	-91.9	-8.4	-66.7	-28.4
ac_da	-30.4	-28.2	-32.4	-21.2
Adf1	-92	-66.9	-52.7	-141.5
Aef1	-249.8	-153.4	-153.9	-316.9
Al	-2.3	Not enriched	-16.5	Not enriched
amos_da	-8.5	-15.6	-9.8	-10.9
Ap	-12.2	Not enriched	-31.6	Not enriched
Ara	-8.8	-3.2	Not enriched	-7.6
ase_da	-18.5	-29	-28.9	-15
ato_da	-4.7	-16.8	-4.1	-15.3
Awh	Not enriched	Not enriched	-11.9	Not enriched
BH1	-21.9	Not enriched	-45	Not enriched
BH2	-8.3	Not enriched	-17.5	Not enriched
Blimp-1	-149	-56.4	-102.3	-172.6
bowl	-2.3	-7.5	-12.9	Not enriched
brk	-5.2	-10.4	Not enriched	-15
Br-Z1	-62.6	-40.7	-33.2	-99.3
Br-Z2	Not enriched	-5.5	Not enriched	-6.8
btd	-73.3	-13.2	-26.3	-60.1
Cad	-97.8	-16.8	-74.9	-41.2
cato_da	Not enriched	-17.5	-1.6	-10.3
Caup	-1.6	Not enriched	Not enriched	Not enriched
CG13897	-37.7	-39.4	-11.3	-102.9
CG33557_da	-71.7	-45.5	-39.4	-72.8
CG3838	-3.4	Not enriched	Not enriched	-4.9
CG4360-F1	-219.8	-141.5	-143.7	-277.4
CG5953	-17.4	-24.6	Not enriched	-92.2
CG8765	-39.7	-21.7	-17.4	-59.6
chinmo	Not enriched	Not enriched	-9.9	Not enriched
ci	-47.8	-51.1	-10.1	-99.6
Crol-F7-16	-174.9	-27.2	-54.2	-146.9
crp	-1.7	-19.5	-8.3	-10.3
D	-5.4	-19	Not enriched	-31.8
da	-7.4	-18.9	-4.8	-15.8
dei_da	Not enriched	-2.9	Not enriched	-2.9
dimm_da	Not enriched	Not enriched	Not enriched	-3.3
Dll	-5.1	Not enriched	-25.9	Not enriched
Dr	-3.1	Not enriched	-25.6	Not enriched
E5	-21.7	Not enriched	-34.4	Not enriched
ecr_usp	Not enriched	Not enriched	-2.1	Not enriched

eg	-4.7	-13.2	-3	-11.6
Ems	-3	Not enriched	-18	Not enriched
En	-18.5	Not enriched	-37.4	Not enriched
Exex	-6.7	Not enriched	-27.4	Not enriched
Fer1_da	-1.3	-16	-3.6	-7.5
Fer3_da	-1.4	Not enriched	-5.3	Not enriched
foxo	-37.5	-41.2	-8.4	-58.9
fru	-14.3	Not enriched	-28.6	Not enriched
HLH4C_da	-46.4	-34.8	-37.4	-32.8
HLH54F_da	-5.8	-10.4	-3.9	-10.7
Hsf	Not enriched	-1.4	Not enriched	Not enriched
inv	Not enriched	Not enriched	-10.8	Not enriched
jim_F1-9	-260	-142	-110.5	-inf
ken	Not enriched	Not enriched	-10.4	Not enriched
kni	-3.1	-9.8	-6.2	-6.4
Kr	-12	Not enriched	Not enriched	Not enriched
l(1)sc_da	-18.9	-14.7	-36.8	-6.2
Lab	-16.1	Not enriched	-33.8	Not enriched
Lag1	Not enriched	-1.7	Not enriched	Not enriched
Lbe	Not enriched	Not enriched	-5.4	Not enriched
Lbl	-3.3	Not enriched	-16.5	Not enriched
Lim1	-17.8	Not enriched	-36	Not enriched
Lola-PD	-95.9	-21.2	-28.3	-79.6
Lola-PF	-8.4	Not enriched	-11.6	Not enriched
Lola-PL	-35.2	-13	-19.8	-22.8
Lola-PO	Not enriched	-12.4	Not enriched	-4
Lola-PQ	-2	-12.8	Not enriched	-35
Mad	-34.3	-26.2	-24.9	-60.3
Med	-23.7	-27.6	-24.4	-51.9
Mirr	-11.8	-2.4	Not enriched	-6.3
nau_da	Not enriched	-3.6	Not enriched	Not enriched
net_da	-7.9	-22.8	-14.6	-13.5
NK7.1	-5.3	Not enriched	-23.1	Not enriched
nub	-10.1	-5.2	-3.3	-15.3
odd	-14.2	-7.5	-30.2	-1.4
opa	-40.6	-9.7	-2.6	-51.8
Otp	-10.3	Not enriched	-28.9	Not enriched
pad	-40.4	-12.6	-13.6	-36.8
Pb	-4.4	Not enriched	-19.9	Not enriched
pdm2	Not enriched	Not enriched	Not enriched	-3.8
pfk	-4.9	Not enriched	-1.9	Not enriched
pnt	Not enriched	-1.8	-2.8	Not enriched
Rel	Not enriched	-6.7	Not enriched	-12.2

retn	-28.3	-14.8	-12.9	-18.7
run_Bgb	-10.6	-5.7	-26.9	-6.8
sage_da	-13.6	-21	-15.9	-15.9
sc_da	-27.6	-25.9	-36.5	-17.8
Scr	-21.2	Not enriched	-37	Not enriched
slp1	-152.7	-169.6	-54	-251.4
sob	-8	-6.4	-21.4	Not enriched
Sox14	Not enriched	-2.7	Not enriched	-1.7
Sox15	Not enriched	-9.2	Not enriched	-17.7
Sp1	-49.4	-11	-11.1	-52.4
sqz	-86.8	-27.4	-22.2	-110.3
sr	-97.1	-36	-28.2	-101
suHw	-4.4	Not enriched	Not enriched	-3.5
tap_da	-12.9	-4.6	-12.4	-6.5
toy	-9.1	Not enriched	Not enriched	-16.2
Trl	-210.2	-83.1	-59.5	-313.6
Ttk-PA	-8.3	Not enriched	Not enriched	-7
Tup	Not enriched	Not enriched	-14.1	Not enriched
Unpg	-5	Not enriched	-24.3	Not enriched
vvl	-8.3	Not enriched	-2.6	-5.6
z	-13.9	Not enriched	Not enriched	-20.9

Table B.6: $log_{10}(padj)$ of the motif enrichment on WD using the list of TFs.

TF	L3IL opening	L3IL closing	WPP opening	WPP closing
AbdA	-34.6	Not enriched	-40	-4.1
ac_da	-27.5	-7.5	-27.7	-18.4
Adf1	-83.7	-44	-78.2	-79.8
Al	-10.6	Not enriched	-11	Not enriched
amos_da	-10.6	-3.6	-13.7	-10.7
Ara	Not enriched	Not enriched	Not enriched	-2.7
ase_da	-22.6	-7.4	-24.2	-15.3
ato_da	-7.3	-6.8	-12.9	-10.4
Awh	-6.3	Not enriched	-7.2	Not enriched
BH1	-29.4	Not enriched	-27.3	Not enriched
BH2	-10.2	Not enriched	-10.7	Not enriched
Blimp-1	-180.2	-35.3	-192.5	-89.6
bowl	-10.7	Not enriched	-9.4	Not enriched
brk	Not enriched	-3.6	Not enriched	-4.9
Br-Z1	-56.4	-20.4	-49.6	-43.5
Cad	-91.8	-15.1	-95.3	-34.8
cato_da	-4.5	-3.4	-9	-4.7
CG13897	-30.7	-41.3	-31.7	-67
CG32105	-66.6	-6.7	-67.7	-19.6
CG33557_da	-65.1	-23.8	-66.8	-48.3
CG3838	Not enriched	Not enriched	Not enriched	-1.4
CG5953	Not enriched	-31.8	Not enriched	-56.8
CG8765	-45.8	-16.8	-45	-38
chinmo	-9.9	Not enriched	-6.7	Not enriched
ci	-28.7	-17.6	-25.1	-54.4
D	Not enriched	-7.8	Not enriched	-15.6
da	-3.1	-4.7	-1.8	-12.5
Deaf1	-29.5	-9.6	-19.8	-15.1
Dll	-18.5	Not enriched	-18.6	Not enriched
eg	-2.6	-3.6	-6.1	-7.6
Ems	-12.1	Not enriched	-15.6	Not enriched
En	-31	Not enriched	-31.9	-3.3
Eve	-4.4	Not enriched	-4.8	Not enriched
Exex	-21.2	Not enriched	-21.5	Not enriched
Fer1_da	-3.1	-1.8	-4	-5.9
Fer3_da	-3.5	Not enriched	-5.5	Not enriched
fru	-33	Not enriched	-34	Not enriched
H2.0	-73.6	-8.8	-76.3	-23
HLH4C_da	-39.2	-11.9	-38.9	-26.2
HLH54F_da	-1.5	-1.4	Not enriched	-8.6
jim_F1-9	-183.2	-119	-163.2	-236.4

klu	-163.8	-70.9	-131.8	-214.2
Kr	-8.9	Not enriched	-4.3	Not enriched
l(1)sc_da	-18.7	Not enriched	-23.5	-4.4
Lab	-28.2	Not enriched	-31.1	-3.4
Lbl	-13.3	Not enriched	-12	Not enriched
Lim1	-30.3	Not enriched	-29.4	-3.2
Lim3	-4.4	Not enriched	-3.6	Not enriched
Lola-PC	-1.3	Not enriched	-7.6	Not enriched
Lola-PD	-62.2	-14.7	-49.6	-37.6
Lola-PF	-8.6	Not enriched	-8.6	Not enriched
Lola-PL	-25.8	-3.5	-30.7	-7
Lola-PO	Not enriched	Not enriched	Not enriched	-6.3
Lola-PQ	Not enriched	-9	Not enriched	-20.5
Mad	-26.7	-11.5	-30.6	-22.3
nau_da	Not enriched	-1.5	Not enriched	Not enriched
net_da	-7.9	-2.8	-10.1	-6.4
NK7.1	-12.2	Not enriched	-18.4	Not enriched
nub	-42.5	-6.2	-30.6	-17.7
odd	-16.6	Not enriched	-19.5	-1.4
opa	-18	-3.7	-10.4	-24.9
Otp	-23.1	Not enriched	-23.4	-1.4
pnt	Not enriched	Not enriched	-5	Not enriched
Repo	-32.1	Not enriched	-32.4	-4
retn	-15.1	-7.1	-14.1	-20
run_Bgb	-47.5	-3.4	-60.9	-8.4
sage_da	-12.5	-4.3	-14.1	-7
sc_da	-28.7	-7.1	-25.9	-16.3
Scr	-34.9	Not enriched	-40.2	-2.4
slp1	-111.6	-64.7	-105.2	-149.9
sna	-7.5	-1.8	-6.7	Not enriched
Sox14	Not enriched	-2	Not enriched	-1.8
Sox15	Not enriched	-4.6	Not enriched	-5.9
sr	-49.8	-24	-40.8	-69.8
suHw	-3.5	Not enriched	-3	-2.1
tap_da	-10.5	Not enriched	-15.5	-3
Tin	-6	Not enriched	-8.2	Not enriched
tj	-9	-2.3	-11.3	-2.2
toy	-5.3	Not enriched	Not enriched	-7.5
Tup	-7.4	Not enriched	-11.5	Not enriched
Ubx	-67.3	-6.5	-70.2	-16.4
Unpg	-16.9	Not enriched	-17.3	Not enriched

Table B.7: $log_{10}(padj)$ **of the motif enrichment on ED using the list of TFs.**

TF	L3IL opening	L3IL closing	WPP opening	WPP closing
AbdA	-55	Not enriched	-22.6	Not enriched
AbdB	-185.8	Not enriched	-88.9	Not enriched
ac_da	-50.9	Not enriched	-35.9	Not enriched
Achi	-inf	Not enriched	Not enriched	Not enriched
Adf1	-225.4	Not enriched	-147.5	-14.3
Aef1	-inf	Not enriched	-inf	-34.1
Al	-9.1	Not enriched	Not enriched	Not enriched
amos_da	-10.9	-2.3	-2.6	-6.2
aop	Not enriched	Not enriched	-6.4	Not enriched
Ap	-26.8	Not enriched	-8.4	Not enriched
Ara	-10.2	Not enriched	-4.5	Not enriched
ase_da	-31.6	Not enriched	-14.9	Not enriched
ato_da	-18.4	-2.6	-6.3	-3.6
bab1	-inf	-11	-inf	-81.1
Bcd	-inf	Not enriched	Not enriched	Not enriched
BH1	-47.6	Not enriched	-21.5	Not enriched
BH2	-20.9	Not enriched	-6.1	Not enriched
Blimp-1	-inf	-15.8	-304.2	-96.9
bowl	-39.8	Not enriched	-43.6	Not enriched
br	-inf	Not enriched	Not enriched	Not enriched
brk	-18.3	-4.7	-13	-6.2
Br-PE	-inf	Not enriched	Not enriched	Not enriched
Br-PL	-inf	Not enriched	Not enriched	Not enriched
Br-Z1	-276.4	-2.9	-173.5	-18
Br-Z4	-inf	Not enriched	Not enriched	Not enriched
btd	-inf	Not enriched	-193.3	-34.9
Cad	-225.1	Not enriched	-108	Not enriched
cato_da	-13.4	Not enriched	-1.6	Not enriched
CG12236	-inf	Not enriched	Not enriched	Not enriched
CG12605	-inf	Not enriched	Not enriched	Not enriched
CG13897	-146.8	-2	-91.1	-18
CG14962	-inf	Not enriched	Not enriched	Not enriched
CG31670	-100.6	Not enriched	-68.1	-24.3
CG32105	-151.9	Not enriched	-70.7	Not enriched
CG33557_da	-152.1	-2.6	-102.8	-20.8
CG3838	-4.2	Not enriched	-7.3	Not enriched
CG4328	-200.8	Not enriched	-96.5	Not enriched
CG4360	-inf	Not enriched	-308.4	-28.1
CG4404	-42.6	Not enriched	-40.9	Not enriched
CG5953	-65.1	-2.1	-25.2	-25.9
CG7368	-inf	-23.2	-inf	-121.9

CG8765	-84	Not enriched	-48.2	-6.3
chinmo	Not enriched	Not enriched	-4.3	Not enriched
ci	-267.5	-5	-156	-56.8
Clk_cyc	-inf	Not enriched	Not enriched	Not enriched
Crc_CG6272	-inf	Not enriched	Not enriched	Not enriched
crol-F7-16	-inf	-6.1	-217.9	-57.9
crp	-4.5	-3	Not enriched	-3
D	-95.4	Not enriched	-63.9	Not enriched
D19A_F10-1	-6	Not enriched	Not enriched	Not enriched
D19B-F10-1	-inf	Not enriched	Not enriched	Not enriched
da	-23	Not enriched	-14.5	-2.1
Deaf1	-60.9	Not enriched	-40.6	Not enriched
Dip3	-inf	Not enriched	Not enriched	Not enriched
Dll	-16.1	Not enriched	-4.3	Not enriched
dm_Max	-inf	Not enriched	Not enriched	Not enriched
Dr	-6.2	Not enriched	Not enriched	Not enriched
E(spl)	-inf	Not enriched	Not enriched	Not enriched
ecr	-inf	Not enriched	Not enriched	Not enriched
eg	-24.7	Not enriched	-13.1	-8.6
Eip75B	-inf	Not enriched	Not enriched	Not enriched
Eip78C	-inf	Not enriched	Not enriched	Not enriched
Ems	-11.9	Not enriched	-3	Not enriched
En	-52.7	Not enriched	-18.4	Not enriched
ERR	-inf	Not enriched	Not enriched	Not enriched
Esg-F3-5	-inf	Not enriched	Not enriched	Not enriched
Ets21c	Not enriched	Not enriched	-2.8	Not enriched
Exd	-inf	Not enriched	Not enriched	Not enriched
Exex	-14.2	Not enriched	-2.8	Not enriched
ey	-144.5	-3	-63.2	-47.7
Fer1	-14.5	Not enriched	Not enriched	Not enriched
Fer3_da	-3.8	Not enriched	Not enriched	Not enriched
fkh	-176.9	Not enriched	-101.8	-10.1
foxo	-108.2	Not enriched	-72.7	-12.1
fru	-19	Not enriched	-19.3	Not enriched
Ftz-f1	-inf	Not enriched	Not enriched	Not enriched
GATAd	-inf	Not enriched	Not enriched	Not enriched
Gsb-n	-inf	Not enriched	Not enriched	Not enriched
gt	-inf	Not enriched	Not enriched	Not enriched
hb	-45	-1.6	-21.6	-21.3
Hbn	-10.7	Not enriched	-1.6	Not enriched
Hey	-14.4	Not enriched	-11.9	Not enriched
Hgtx	-73.4	Not enriched	-39.4	Not enriched
HLH4C	-31.2	Not enriched	-15.8	Not enriched

HLH54F_da	-2.5	Not enriched	-1.6	Not enriched
HLHm3	-inf	Not enriched	Not enriched	Not enriched
HLHmbeta	-inf	Not enriched	Not enriched	Not enriched
HLHmdelta	-inf	Not enriched	Not enriched	Not enriched
HLHmgamma	-inf	Not enriched	Not enriched	Not enriched
Hnf4	-inf	Not enriched	Not enriched	Not enriched
Hr46	-inf	Not enriched	Not enriched	Not enriched
Hr78	-inf	Not enriched	Not enriched	Not enriched
Hth	-inf	Not enriched	Not enriched	Not enriched
jim_F1-9	-inf	-4.9	-291.2	-56.8
kay_Jra	-inf	Not enriched	Not enriched	Not enriched
ken	-23.3	Not enriched	-17.7	-6.9
klu	-inf	-12	-inf	-156.2
Kr	-57.3	Not enriched	-43.7	-7
l(1)sc_da	-13.1	Not enriched	-11.9	Not enriched
Lab	-33.4	Not enriched	-11.8	Not enriched
Lbl	-2.3	Not enriched	Not enriched	Not enriched
Lim1	-43.5	Not enriched	-15.3	Not enriched
lola	-inf	Not enriched	Not enriched	Not enriched
Lola-PA	-inf	Not enriched	Not enriched	Not enriched
Lola-PC	-150.9	-1.7	-115.4	-4.9
Lola-PD	-77.5	-3	-31.3	-28.6
Lola-PF	-18.4	Not enriched	-12.2	Not enriched
Lola-PG	-inf	Not enriched	Not enriched	Not enriched
Lola-PK	-inf	Not enriched	Not enriched	Not enriched
Lola-PL	-76.2	-10.2	-43.4	-27.8
Lola-PQ	-34.7	Not enriched	-18.8	Not enriched
Lola-PT	-inf	Not enriched	Not enriched	Not enriched
Lola-PU	-inf	Not enriched	Not enriched	Not enriched
luna	-133.6	-3.4	-71.1	-24.1
Mad	-111.3	-1.9	-89.4	-2.7
Max_Mnt	-inf	Not enriched	Not enriched	Not enriched
Mio_bigmax	-inf	Not enriched	Not enriched	Not enriched
Mirr	-9.9	Not enriched	-5.8	Not enriched
net_da	-17.7	-2.8	-9.4	-3.5
NK7.1	-19.8	Not enriched	-7	Not enriched
nub	-133.4	Not enriched	-95.8	-9.7
Oc	-inf	Not enriched	Not enriched	Not enriched
odd	-31.8	Not enriched	-35.2	Not enriched
opa	-167.8	-5.2	-97.4	-34.4
Optix	-inf	Not enriched	Not enriched	Not enriched
Otp	-23.8	Not enriched	-6.2	Not enriched
ovo	-inf	Not enriched	Not enriched	Not enriched

pad	-129.2	Not enriched	-97.4	-7.1
pan	-inf	Not enriched	Not enriched	Not enriched
Pb	-7.7	Not enriched	Not enriched	Not enriched
pfk	-22.4	Not enriched	-6.1	Not enriched
PhdP	-44.1	Not enriched	-14.4	Not enriched
pho	-12.8	-1.7	-7.9	Not enriched
pnt	-4.7	Not enriched	-6.7	Not enriched
Poxn	-inf	Not enriched	Not enriched	Not enriched
Ptx1	-inf	Not enriched	Not enriched	Not enriched
Rel	-19.6	Not enriched	-13.7	-4.4
Repo	-55.8	Not enriched	-21.5	Not enriched
retn	-69.6	Not enriched	-37.3	Not enriched
run_Bgb	-13.3	Not enriched	-5.8	Not enriched
sage_da	-21.6	Not enriched	-15.3	-1.6
sc_da	-42.3	-2	-33	Not enriched
Scr	-43.2	Not enriched	-17.1	Not enriched
scrt	-inf	Not enriched	Not enriched	Not enriched
sens	-inf	Not enriched	Not enriched	Not enriched
Six4	-inf	Not enriched	Not enriched	Not enriched
Slou	-34	Not enriched	-12.4	Not enriched
slp1	-inf	-2.4	-313.4	-51.7
slp2	-220.9	Not enriched	-124.5	-8
sob	-25.3	Not enriched	-31.7	Not enriched
Sox14	-31	Not enriched	-22.4	Not enriched
Sp1	-287.8	-2.4	-144.5	-34.8
sr	-inf	-5.2	-261.4	-67.4
ss_tgo	-inf	Not enriched	Not enriched	Not enriched
SuH	-inf	Not enriched	Not enriched	Not enriched
sv	-inf	Not enriched	Not enriched	Not enriched
svp	-inf	Not enriched	Not enriched	Not enriched
tap_da	-35.2	Not enriched	-27.9	-2
tgo_cyc	-inf	Not enriched	Not enriched	Not enriched
tgo_sim	-inf	Not enriched	Not enriched	Not enriched
tgo_ss	-inf	Not enriched	Not enriched	Not enriched
tgo_tai	-inf	Not enriched	Not enriched	Not enriched
Tin	Not enriched	Not enriched	-2.4	Not enriched
tj	-15.8	Not enriched	-10	Not enriched
tll	-23.7	Not enriched	-8.6	Not enriched
toy	-25.8	-1.3	-8.7	-20.7
Trl	-inf	-9.4	-228.8	-120.9
ttk	-3.1	Not enriched	-4.7	Not enriched
Ttk-PA	-102.1	-1.6	-66.9	-24
Ttk-PF	-17.3	Not enriched	-20.4	Not enriched

Tup	-9	Not enriched	Not enriched	Not enriched
twi_da	-inf	Not enriched	Not enriched	Not enriched
Ubx	-128.7	Not enriched	-60.1	Not enriched
Unpg	-14.2	Not enriched	-2.1	Not enriched
Usf	-inf	Not enriched	Not enriched	Not enriched
vfl	-inf	Not enriched	Not enriched	Not enriched
vri	-inf	Not enriched	Not enriched	Not enriched
vvl	-92.4	Not enriched	-84.3	-2.6
wor	-inf	Not enriched	Not enriched	Not enriched
Xrp1_CG627	-inf	Not enriched	Not enriched	Not enriched

Table B.8: $log_{10}(padj)$ **of the motif enrichment on CNS using the list of TFs.**

TF	L3IL opening	L3IL closing	WPP opening	WPP closing
AbdA	Not enriched	Not enriched	-13.3	Not enriched
AbdB	-8.3	Not enriched	-58.2	-5.5
ac_da	-72.4	-47.4	-149	-110.9
Adf1	-54.4	-39.7	-195	-40.3
Aef1	-129.3	-71.7	-inf	-61.3
amos_da	-41.6	-39.1	-72.9	-70.2
aop	-2.1	Not enriched	Not enriched	Not enriched
Ara	-1.8	-4.8	-51.4	Not enriched
ase_da	-56.1	-46	-115	-102.9
ato_da	-45	-43.9	-74.9	-89.7
BH1	Not enriched	Not enriched	-3.9	Not enriched
bin	-42.2	-43.4	-152.6	-43.1
Blimp-1	-135.5	-29.7	-inf	-55.2
bowl	-1.3	Not enriched	-13	-3
brk	-12.6	-1.5	-31.6	-3.1
Br-Z1	-26	-16.6	-76.5	-20.6
btd	-60.6	-17.1	-121.9	-27.6
Cad	-15.2	-4.4	-91.8	-11.9
cato_da	-34	-45	-86.1	-85
Caup	Not enriched	Not enriched	-5.5	Not enriched
CG10267	-20.3	-6.6	-108.8	Not enriched
CG13897	-27.8	-27.9	-108.3	-33.1
CG31670	-21.7	-18.2	-83.7	-7.4
CG33557_da	-56	-36.3	-132	-72.9
CG3838	Not enriched	Not enriched	-3.6	Not enriched
CG4360	-108.7	-61.1	-inf	-49.7
CG5669	-28.1	-8.2	-54.2	-21
CG5953	-19.6	-16.3	-23.5	-15.3
CG6276	Not enriched	Not enriched	-2.7	Not enriched
CG8765	-10.4	-9.6	-59.2	-5
ci	-81.5	-38.1	-174.5	-53.9
crol-F7-16	-92.3	-31.6	-232.7	-36.3
crp	-32.4	-40.6	-92.8	-66.4
D	-2.6	-7.3	-209.9	Not enriched
da	-55.3	-39.4	-84.1	-81.4
Deaf1	Not enriched	-7.4	-5.5	-12.4
dei_da	-22.9	-21.6	-18	-36
dimm_da	-13.8	-11.7	-9.2	-13.9
E5	Not enriched	Not enriched	-4.3	Not enriched
eg	-11.6	-10.3	-19.3	-5.3
En	Not enriched	Not enriched	-12.6	Not enriched

Esg-F3-5	Not enriched	Not enriched	Not enriched	-1.8
Ets97D	-10.6	Not enriched	-2.7	Not enriched
ey	-42	-8.8	-65.8	-10.5
Fer1_da	-48.6	-49.2	-99.4	-102
Fer2_da	-7.4	-16.6	-26.9	-38.3
Fer3_da	-15.8	-29.4	-47.2	-66.9
fkh	-29.6	-98.1	-104.2	-95.1
foxo	-46.9	-32.9	-137	-31.3
fru	Not enriched	Not enriched	-11.7	Not enriched
grh	Not enriched	Not enriched	-7.3	Not enriched
Hand_da	-2.7	-5.7	-2	-12.7
hb	-21.7	-18	-56.4	-12.2
HLH4C_da	-75.3	-51.4	-158.5	-114.5
HLH54F	-38.1	-23.9	-74.9	-53.2
jim_F1-9	-147.2	-82.1	-inf	-64.5
ken	-2.1	Not enriched	-9.3	Not enriched
klu	-211.1	-78.5	-inf	-107.5
kni	-5.6	-8.3	-19.5	-5.4
l(1)sc_da	-42.2	-18.6	-86.6	-58.4
Lab	Not enriched	Not enriched	-5.6	Not enriched
Lim1	Not enriched	Not enriched	-8.6	Not enriched
Lola-PC	-25	Not enriched	-41.6	-6.9
Lola-PD	-56.5	-25.1	-120.1	-30.4
Lola-PF	Not enriched	Not enriched	Not enriched	-2.5
Lola-PL	-40.9	-17.1	-105	-33.8
Lola-PO	Not enriched	-4.8	-4.8	-5.3
Lola-PQ	-1.7	-9.8	-81.5	Not enriched
luna	-32.9	-4.2	-46.5	-18.3
Mad	-19.9	-14.6	-86.8	-18.8
Max_Mnt	Not enriched	Not enriched	Not enriched	-2.9
Med	-15.7	-8.1	-69.7	-17.3
Met_Clk	Not enriched	Not enriched	Not enriched	-3.6
Mirr	-2.3	-5.9	-56.8	Not enriched
nau_da	-20.1	-19.3	-29.3	-58.2
net_da	-68.8	-63.3	-139.6	-124.6
NFAT	-8.7	-1.4	-30.6	-2.7
nub	-10.5	-7	-19.8	-16.4
odd	-4.3	Not enriched	-20.9	Not enriched
Oli_da	Not enriched	Not enriched	Not enriched	-5.3
opa	-52.3	-13.5	-123.8	-26.1
Otp	Not enriched	Not enriched	-1.7	Not enriched
pad	-21	-4.2	-56	-7.5
pdm2	-1.8	-2.4	-3.7	-8.6

pnr	-2.5	-5.5	-28	-2.2
pnt	-1.4	Not enriched	Not enriched	Not enriched
Rel	-10.4	-1.5	-18	-6.5
Repo	Not enriched	Not enriched	-12.1	Not enriched
run_Bgb	Not enriched	Not enriched	-9.6	Not enriched
sage_da	-56	-60.6	-116.4	-110
sc_da	-71.9	-53.6	-134.7	-118.5
Scr	Not enriched	Not enriched	-1.8	Not enriched
slp1	-134.2	-112.9	-inf	-114.1
slp2	-53.3	-67.6	-207.3	-57.8
sob	Not enriched	-2.2	-6.6	-2.2
Sox14	Not enriched	Not enriched	-40.9	Not enriched
Sox15	-3.8	-10.3	-168.7	Not enriched
Sp1	-55.9	-11.3	-92.7	-32.3
sqz	-43	-30.1	-103.7	-23.5
sr	-98.3	-40.4	-221	-55.8
sug	-35.4	-10	-72.8	-18
suHw	-8.7	Not enriched	-4.1	-1.5
tai_Clk	Not enriched	Not enriched	Not enriched	-3.2
tap_da	-23.7	-24.8	-40.5	-49.2
toy	-18.2	-2.1	-38	-1.8
Trl	-147.2	-67.5	-inf	-53.7
Ttk-PA	-11.5	Not enriched	-12.3	Not enriched
twi_da	Not enriched	-5.8	Not enriched	-19.1
Ubx	-3.7	Not enriched	-40	-2.8
vvl	-1.3	Not enriched	-7.1	Not enriched
wor	Not enriched	Not enriched	Not enriched	-7.6
z	-19.7	-6.8	-30.7	-15.3

Table B.9: $log_{10}(padj)$ **of the motif enrichment on SG using the list of TFs.**

achi	CG7745	her	NFAT
Adf1	CG8281	HHEX	NK7.1
Aef1	CG8319	HLH106	ovo
aop	CG8765	HLHmbeta	pad
ap	chinmo	Hmx	pan
Asciz	crc	Hnf4	pfk
Atf-2	CrebA	hng1	pho
Atf6	crol	hng3	phol
BEAF-32	crp	Hr39	pnr
bigmax	ct	Hr4	pnt
bowl	CTCF	Hr46	Rel
br-Z1	cwo	Hr51	run
br-Z2	cyc	Hr78	sage
br-Z3	D19A	Hsf	schlank
br-Z4	D19B	Irbp18	sd
brk	da	jigr1	shn
BtbVII	Deaf1	jim	sima
Btn	dl	Jra	Sox14
Cf2	Dref	Kah	Spps
CG10904	E(spl)mbeta-HLH	kay	sqz
CG11504	EcR	ken	srp
CG12155	EcR-usp	l(3)neo38	Stat92E
CG12236	Eip74EF	Lag1	Su(H)
CG12768	Eip75B	lola-PC	su(Hw)
CG15601	Eip78C	lola-PO	sug
CG3065	Eip93F	lola-PY	tai
CG34031	ERR	luna	Trl
CG3407	Ets21c	Mad	ttk
CG3838	Ets97D	Max	twi
CG3919	Exd	Med	Usf
CG4360	foxo	Mes2	usp
CG4404	fru	Met	vfl
CG4854	ftz-f1	Mitf	vri
CG5180	GATAd	Mnt	Xrp1
CG5953	GATAe	Mondo	z
CG6276	gsb	nau	Zif
CG7386	h	net	ZIPIC

Table B.10: List of expressed TFs in the S2 cells ecdysone response used for the modeling.

Assign.	Expr.	Genes	Acc.	Dec.	Window	Mean MSE	Mean PCC
Window	Yes	All	No	Yes	50000	0.835	0.396
Window	No	All	No	Yes	50000	0.841	0.399
Window	Yes	All	Peaks	Yes	50000	0.846	0.391
Window	Yes	All	Full	Yes	50000	0.855	0.377
Window	Yes	All	No	Yes	20000	0.864	0.380
Window	Yes	All	No	Yes	10000	0.867	0.355
Window	Yes	All	Peaks	Yes	20000	0.869	0.362
Window	No	All	Peaks	Yes	50000	0.871	0.364
Nearest	Yes	Diff	No	Yes	Not defined	0.872	0.371
Nearest	Yes	Diff	Peaks	Yes	Not defined	0.875	0.349
Nearest	Yes	Diff	Full	Yes	Not defined	0.878	0.352
Window	Yes	All	Full	Yes	20000	0.881	0.351
Window	No	All	No	Yes	20000	0.882	0.364
Window	No	All	Full	Yes	50000	0.882	0.347
Window	Yes	All	Full	Yes	10000	0.886	0.327
Nearest	No	Diff	No	Yes	Not defined	0.886	0.328
Window	Yes	All	Peaks	Yes	10000	0.886	0.336
Window	No	All	No	Yes	10000	0.887	0.346
Nearest	No	Diff	Peaks	Yes	Not defined	0.888	0.321
Window	No	All	Peaks	Yes	20000	0.888	0.327
Window	Yes	All	No	No	10000	0.888	0.326
Window	Yes	All	Peaks	No	20000	0.892	0.334
Window	Yes	All	No	Yes	5000	0.893	0.310
Regions	Yes	All	Peaks	Yes	Not defined	0.893	0.318
Regions	Yes	Diff	Full	Yes	Not defined	0.894	0.313
Window	Yes	All	Full	No	5000	0.895	0.297
Window	Yes	All	No	No	20000	0.897	0.335
Regions	Yes	Diff	No	Yes	Not defined	0.897	0.323
Regions	No	Diff	No	Yes	Not defined	0.897	0.307
Window	No	All	No	No	10000	0.897	0.336
Regions	Yes	Expr	Full	Yes	Not defined	0.901	0.300
Window	No	All	Full	Yes	20000	0.902	0.337
Window	Yes	All	No	No	50000	0.904	0.316
Window	No	All	No	No	50000	0.904	0.304
Regions	Yes	Expr	No	Yes	Not defined	0.906	0.304
Window	No	All	No	No	20000	0.907	0.320
Window	Yes	All	No	No	5000	0.908	0.318
Window	Yes	All	Peaks	No	10000	0.909	0.320
Window	Yes	All	Peaks	Yes	5000	0.910	0.313
Regions	Yes	Diff	Peaks	Yes	Not defined	0.910	0.317
Nearest	No	Diff	Full	Yes	Not defined	0.910	0.332

Regions	No	Diff	Full	Yes	Not defined	0.910	0.287
Window	No	All	Full	Yes	10000	0.911	0.285
Regions	No	All	Full	Yes	Not defined	0.911	0.292
Regions	Yes	Expr	Peaks	Yes	Not defined	0.911	0.302
Window	No	All	Peaks	Yes	10000	0.912	0.296
Window	No	All	Peaks	No	20000	0.912	0.305
Regions	Yes	All	Full	Yes	Not defined	0.912	0.329
Window	Yes	All	Full	No	20000	0.913	0.308
Window	Yes	All	Full	Yes	5000	0.913	0.296
Regions	Yes	All	No	Yes	Not defined	0.913	0.321
Window	Yes	All	Full	No	10000	0.915	0.320
Window	Yes	All	Peaks	No	5000	0.915	0.309
Window	No	All	Full	No	20000	0.917	0.295
Window	Yes	All	Peaks	No	50000	0.918	0.302
Regions	No	All	Peaks	Yes	Not defined	0.918	0.285
Window	No	All	Full	No	10000	0.919	0.282
Window	Yes	All	Full	No	50000	0.919	0.309
Nearest	Yes	All	Peaks	Yes	Not defined	0.920	0.292
Regions	No	All	No	Yes	Not defined	0.921	0.293
Window	No	All	No	Yes	5000	0.924	0.290
Window	No	All	Full	No	50000	0.924	0.282
Regions	No	Expr	Peaks	Yes	Not defined	0.926	0.276
Nearest	Yes	Expr	Full	Yes	Not defined	0.926	0.258
Regions	No	Diff	Peaks	Yes	Not defined	0.927	0.281
Regions	No	Expr	No	Yes	Not defined	0.927	0.299
Regions	Yes	All	Full	No	Not defined	0.927	0.242
Nearest	Yes	All	Full	Yes	Not defined	0.932	0.275
Regions	No	Expr	Full	Yes	Not defined	0.933	0.265
Window	No	All	Peaks	Yes	5000	0.934	0.241
Window	No	All	Peaks	No	50000	0.935	0.286
Window	No	All	Peaks	No	10000	0.935	0.289
Nearest	Yes	Expr	No	Yes	Not defined	0.936	0.265
Nearest	Yes	All	No	Yes	Not defined	0.936	0.253
Nearest	Yes	All	Peaks	No	Not defined	0.936	0.249
Nearest	No	All	No	Yes	Not defined	0.937	0.250
Regions	Yes	Diff	Peaks	No	Not defined	0.937	0.247
Window	No	All	Peaks	No	5000	0.937	0.270
Nearest	Yes	All	Full	No	Not defined	0.938	0.246
Regions	Yes	All	Peaks	No	Not defined	0.939	0.249
Window	No	All	No	No	5000	0.940	0.270
Nearest	No	Expr	Full	Yes	Not defined	0.941	0.229
Nearest	No	Expr	No	Yes	Not defined	0.943	0.244
Regions	Yes	Diff	Full	No	Not defined	0.944	0.257

Regions	Yes	Expr	Full	No	Not defined	0.944	0.246
Regions	Yes	Diff	No	No	Not defined	0.945	0.246
Nearest	No	All	Full	Yes	Not defined	0.945	0.240
Regions	No	Expr	Peaks	No	Not defined	0.947	0.222
Nearest	No	All	Full	No	Not defined	0.947	0.202
Regions	No	Diff	Peaks	No	Not defined	0.947	0.220
Window	No	All	Full	Yes	5000	0.947	0.253
Nearest	No	Expr	Peaks	Yes	Not defined	0.948	0.231
Regions	No	Diff	Full	No	Not defined	0.949	0.227
Regions	Yes	Expr	No	No	Not defined	0.949	0.241
Regions	No	All	No	No	Not defined	0.950	0.231
Nearest	Yes	Expr	Peaks	Yes	Not defined	0.951	0.255
Regions	Yes	All	No	No	Not defined	0.951	0.250
Regions	Yes	Expr	Peaks	No	Not defined	0.952	0.237
Window	No	All	Full	No	5000	0.952	0.261
Regions	No	All	Full	No	Not defined	0.955	0.231
Nearest	Yes	All	No	No	Not defined	0.956	0.235
Regions	No	All	Peaks	No	Not defined	0.959	0.226
Regions	No	Expr	Full	No	Not defined	0.961	0.225
Nearest	No	All	Peaks	No	Not defined	0.961	0.212
Window	Yes	All	Peaks	No	2000	0.962	0.186
Window	Yes	All	No	No	2000	0.962	0.213
Window	Yes	All	Full	No	2000	0.964	0.187
Window	Yes	All	No	Yes	2000	0.965	0.174
Nearest	Yes	Expr	Peaks	No	Not defined	0.967	0.179
Regions	No	Diff	No	No	Not defined	0.968	0.221
Nearest	Yes	Expr	Full	No	Not defined	0.969	0.170
Nearest	No	All	No	No	Not defined	0.969	0.199
Nearest	No	All	Peaks	Yes	Not defined	0.969	0.228
Window	Yes	All	Full	Yes	2000	0.971	0.195
Regions	No	Expr	No	No	Not defined	0.971	0.215
Window	Yes	All	Peaks	Yes	2000	0.974	0.188
Nearest	Yes	Expr	No	No	Not defined	0.975	0.165
Window	Yes	All	Peaks	No	1000	0.976	0.134
Nearest	No	Expr	No	No	Not defined	0.978	0.169
Window	No	All	Full	Yes	1000	0.979	0.098
Nearest	No	Expr	Peaks	No	Not defined	0.980	0.154
Window	Yes	All	Peaks	Yes	1000	0.980	0.138
Window	Yes	All	No	No	1000	0.981	0.143
Nearest	No	Diff	No	No	Not defined	0.986	0.133
Nearest	Yes	Diff	No	No	Not defined	0.987	0.141
Nearest	Yes	Diff	Peaks	No	Not defined	0.988	0.150
Nearest	No	Diff	Full	No	Not defined	0.988	0.140

Nearest	Yes	Diff	Full	No	Not defined	0.989	0.143
Nearest	No	Diff	Peaks	No	Not defined	0.991	0.126
Nearest	No	Expr	Full	No	Not defined	0.994	0.132
Window	Yes	All	Full	No	1000	0.995	0.128
Window	Yes	All	No	Yes	1000	1.000	0.135
Window	No	All	Full	Yes	2000	1.005	0.137
Window	No	All	No	No	2000	1.009	0.162
Window	No	All	Peaks	Yes	2000	1.009	0.140
Window	No	All	Full	No	2000	1.009	0.126
Window	No	All	Peaks	No	1000	1.016	0.106
Window	No	All	Peaks	No	2000	1.033	0.144
Window	Yes	All	Full	Yes	1000	1.039	0.134
Window	No	All	Full	No	1000	1.046	0.103
Window	No	All	No	Yes	2000	1.080	0.164
Window	No	All	No	Yes	1000	1.109	0.103
Window	No	All	Peaks	Yes	1000	1.168	0.099
Window	No	All	No	No	1000	2.828	0.107

Table B.11: List of performances for every definition of TF-gene scores.

Appendix C

Abbreviations

Abbreviations

bp base pairs

CNS central nervous system

DHSs DNase I hypersensitive sites

E3IL early third instar larva

ED eye disc

FC fold change

FPKM fragments per kilobase of transcript per million mapped reads

GO gene ontology

GTFs general transcription factors

L3IL late third instar larva

MSE mean squared error

PCA principal component analysis

PCC Pearson correlation coefficient

PWM position weight matrix

PWMs position weight matrices

SG salivary glands

TF transcription factor

TFBS transcription factor binding site

TFBSs transcription factor binding sites

TFs transcription factors

Tn-D transition down

Tn-U transition up

TSS transcription start site

TSSs transcription start sites

Tt-D transient down

Tt-U transient up

UTC untreated control

WD wing disc

WPP white prepupa

List of Figures

List of Tables

Bibliography

[Arnold et al., 2013] Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, Ł. M., Rath, M., and Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by starr-seq. *Science*, 339(6123):1074–1077.

[Ashburner et al., 2000] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25.

[Badis et al., 2009] Badis, G., Berger, M. F., Philippakis, A. A., Talukder, S., Gehrke, A. R., Jaeger, S. A., Chan, E. T., Metzler, G., Vedenko, A., Chen, X., et al. (2009). Diversity and complexity in dna recognition by transcription factors. *Science*, 324(5935):1720–1723.

[Baehrecke, 1996] Baehrecke, E. H. (1996). Ecdysone signaling cascade and regulation of drosophila metamorphosis. *Archives of insect biochemistry and physiology*, 33(3-4):231–244.

[Bailey et al., 2009] Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., and Noble, W. S. (2009). Meme suite: tools for motif discovery and searching. *Nucleic acids research*, 37(suppl_2):W202–W208.

[Blatti et al., 2015] Blatti, C., Kazemian, M., Wolfe, S., Brodsky, M., and Sinha, S. (2015). Integrating motif, dna accessibility and gene expression data to build regulatory maps in an organism. *Nucleic acids research*, 43(8):3998–4012.

[Boyle et al., 2008] Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S., and Crawford, G. E. (2008). High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–322.

[Buenrostro et al., 2013] Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nature methods*, 10(12):1213.

[Burtis et al., 1990] Burtis, K. C., Thummel, C. S., Jones, C. W., Karim, F. D., and Hogness, D. S. (1990). The drosophila 74ef early puff contains e74, a complex ecdysone-inducible gene that encodes two ets-related proteins. *Cell*, 61(1):85–99.

[Chao and Guild, 1986] Chao, A. T. and Guild, G. M. (1986). Molecular analysis of the ecdysterone-inducible 2b5 'early' puff in drosophila melanogaster. *The EMBO journal*, 5(1):143–150.

[Chechik and Koller, 2009] Chechik, G. and Koller, D. (2009). Timing of gene expression responses to environmental changes. *Journal of Computational Biology*, 16(2):279–290.

[Consortium, 2016] Consortium, G. O. (2016). Expansion of the gene ontology knowledgebase and resources. *Nucleic acids research*, 45(D1):D331–D338.

[Cuellar-Partida et al., 2011] Cuellar-Partida, G., Buske, F. A., McLeay, R. C., Whitington, T., Noble, W. S., and Bailey, T. L. (2011). Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics*, 28(1):56–62.

[Durek et al., 2016] Durek, P., Nordström, K., Gasparoni, G., Salhab, A., Kressler, C., De Almeida, M., Bassler, K., Ulas, T., Schmidt, F., Xiong, J., et al. (2016). Epigenomic profiling of human cd4+ t cells supports a linear differentiation model and highlights molecular regulators of memory development. *Immunity*, 45(5):1148–1161.

[Duren et al., 2017] Duren, Z., Chen, X., Jiang, R., Wang, Y., and Wong, W. H. (2017). Modeling gene regulation from paired expression and chromatin accessibility data. *Proceedings of the National Academy of Sciences*, 114(25):E4914–E4923.

[Durinck et al., 2009] Durinck, S., Spellman, P. T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nature protocols*, 4(8):1184.

[Emery et al., 1994] Emery, I. F., Bedian, V., and Guild, G. M. (1994). Differential expression of broad-complex transcription factors may forecast tissue-specific developmental fates during drosophila metamorphosis. *Development*, 120(11):3275–3287.

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

[Fischer et al., 2017] Fischer, D. S., Theis, F. J., and Yosef, N. (2017). Impulse model-based differential expression analysis of time course sequencing data. *bioRxiv*, page 113548.

[Friedman et al., 2010] Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

[Frühauf, 2015] Frühauf, K. (2015). Dissecting the regulation of gene expression during steroid hormone signaling in drosophila by dynamic transcriptome analysis (dta).

[Gauhar et al., 2009] Gauhar, Z., Sun, L. V., Hua, S., Mason, C. E., Fuchs, F., Li, T.-R., Boutros, M., and White, K. P. (2009). Genomic mapping of binding regions for the ecdysone receptor protein complex. *Genome research*, 19(6):1006–1013.

[Giardine et al., 2005] Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., et al. (2005). Galaxy: a platform for interactive large-scale genome analysis. *Genome research*, 15(10):1451–1455.

[Gibert et al., 2005] Gibert, J.-M., Marcellini, S., David, J. R., Schlötterer, C., and Simpson, P. (2005). A major bristle qtl from a selected population of drosophila uncovers the zinc-finger transcription factor poils-au-dos, a repressor of achaete–scute. *Developmental biology*, 288(1):194–205.

[Goulev et al., 2008] Goulev, Y., Fauny, J. D., Gonzalez-Marti, B., Flagiello, D., Silber, J., and Zider, A. (2008). Scalloped interacts with yorkie, the nuclear effector of the hippo tumor-suppressor pathway in drosophila. *Current Biology*, 18(6):435–441.

[Gramates et al., 2016] Gramates, L. S., Marygold, S. J., Santos, G. d., Urbano, J.-M., Antonazzo, G., Matthews, B. B., Rey, A. J., Tabone, C. J., Crosby, M. A., Emmert, D. B., et al. (2016). Flybase at 25: looking to the future. *Nucleic acids research*, page gkw1016.

[Grant et al., 2011] Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). Fimo: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018.

[Gusmao et al., 2016] Gusmao, E. G., Allhoff, M., Zenke, M., and Costa, I. G. (2016). Analysis of computational footprinting methods for dnase sequencing experiments. *Nature methods*, 13(4):303.

[Gusmao et al., 2014] Gusmao, E. G., Dieterich, C., Zenke, M., and Costa, I. G. (2014). Detection of active transcription factor binding sites with the combination of dnase hypersensitivity and histone modifications. *Bioinformatics*, 30(22):3143–3151.

[He et al., 2014] He, H. H., Meyer, C. A., Chen, M.-W., Zang, C., Liu, Y., Rao, P. K., Fei, T., Xu, H., Long, H., Liu, X. S., et al. (2014). Refined dnase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nature methods*, 11(1):73.

[Hesselberth et al., 2009] Hesselberth, J. R., Chen, X., Zhang, Z., Sabo, P. J., Sandstrom, R., Reynolds, A. P., Thurman, R. E., Neph, S., Kuehn, M. S., Noble, W. S., et al. (2009). Global mapping of protein-dna interactions in vivo by digital genomic footprinting. *Nature methods*, 6(4):283.

[Hill et al., 2013] Hill, R. J., Billas, I. M., Bonneton, F., Graham, L. D., and Lawrence, M. C. (2013). Ecdysone receptors: from the ashburner model to structural biology. *Annual review of entomology*, 58:251–271.

[Hiruma and Riddiford, 2004] Hiruma, K. and Riddiford, L. M. (2004). Differential control of mhr3 promoter activity by isoforms of the ecdysone receptor and inhibitory effects of e75a and mhr3. *Developmental biology*, 272(2):510–521.

[Horner et al., 1995] Horner, M. A., Chen, T., and Thummel, C. S. (1995). Ecdysteroid regulation and dna binding properties of drosophila nuclear hormone receptor superfamily members. *Developmental biology*, 168(2):490–502.

[Huber et al., 2015] Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., et al. (2015). Orchestrating high-throughput genomic analysis with bioconductor. *Nature methods*, 12(2):115.

[Huet et al., 1995] Huet, F., Ruiz, C., and Richards, G. (1995). Sequential gene activation by ecdysone in drosophila melanogaster: the hierarchical equivalence of early and early late genes. *Development*, 121(4):1195–1204.

[Jaccard, 1901] Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579.

[Johnson et al., 2007] Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–1502.

[Jolma et al., 2013] Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., et al. (2013). Dna-binding specificities of human transcription factors. *Cell*, 152(1-2):327–339.

[Jung et al., 2018] Jung, C., Bandilla, P., von Reutern, M., Schnepf, M., Rieder, S., Unnerstall, U., and Gaul, U. (2018). True equilibrium measurement of transcription factor-dna binding affinities using automated polarization microscopy. *Nature communications*, 9(1):1605.

[Juven-Gershon et al., 2008] Juven-Gershon, T., Hsu, J.-Y., and Kadonaga, J. T. (2008). Caudal, a key developmental regulator, is a dpe-specific transcriptional factor. *Genes & development*, 22(20):2823–2830.

[Kalm et al., 1994] Kalm, v. L., Crossgrove, K., Von Seggern, D., Guild, G. M., and Beckendorf, S. K. (1994). The broad-complex directly controls a tissue-specific response to the steroid hormone ecdysone at the onset of drosophila metamorphosis. *The EMBO Journal*, 13(15):3505–3516.

[Kaplan et al., 2011] Kaplan, T., Li, X.-Y., Sabo, P. J., Thomas, S., Stamatoyannopoulos, J. A., Biggin, M. D., and Eisen, M. B. (2011). Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early drosophila development. *PLoS genetics*, 7(2):e1001290.

[Karim et al., 1993] Karim, F. D., Guild, G. M., and Thummel, C. S. (1993). The drosophila broad-complex plays a key role in controlling ecdysone-regulated gene expression at the onset of metamorphosis. *Development*, 118(3):977–988.

[Karim and Thummel, 1992] Karim, F. D. and Thummel, C. S. (1992). Temporal coordination of regulatory gene expression by the steroid hormone ecdysone. *The EMBO Journal*, 11(11):4083–4093.

[Kharchenko et al., 2011] Kharchenko, P. V., Alekseyenko, A. A., Schwartz, Y. B., Minoda, A., Riddle, N. C., Ernst, J., Sabo, P. J., Larschan, E., Gorchakov, A. A., Gu, T., et al. (2011). Comprehensive analysis of the chromatin landscape in drosophila melanogaster. *Nature*, 471(7339):480.

[Koelle et al., 1991] Koelle, M. R., Talbot, W. S., Segraves, W. A., Bender, M. T., Cherbas, P., and Hogness, D. S. (1991). The drosophila ecr gene encodes an ecdysone receptor, a new member of the steroid receptor superfamily. *Cell*, 67(1):59–77.

[Koohy et al., 2013] Koohy, H., Down, T. A., and Hubbard, T. J. (2013). Chromatin accessibility data sets show bias due to sequence specificity of the dnase i enzyme. *PloS one*, 8(7):e69853.

[Koyama et al., 2014] Koyama, T., Rodrigues, M. A., Athanasiadis, A., Shingleton, A. W., and Mirth, C. K. (2014). Nutritional control of body size through foxo-ultraspiracle mediated ecdysone biosynthesis. *Elife*, 3.

[Kvon et al., 2014] Kvon, E. Z., Kazmar, T., Stampfel, G., Yáñez-Cuna, J. O., Pagani, M., Schernhuber, K., Dickson, B. J., and Stark, A. (2014). Genome-scale functional characterization of drosophila developmental enhancers in vivo. *Nature*, 512(7512):91.

[Langmead and Salzberg, 2012] Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357.

[Levine, 2010] Levine, M. (2010). Transcriptional enhancers in animal development and evolution. *Current Biology*, 20(17):R754–R763.

[Li et al., 2009] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079.

[Li and Liberles, 2015] Li, Q. and Liberles, S. D. (2015). Aversion and attraction through olfaction. *Current Biology*, 25(3):R120–R129.

[Li and White, 2003] Li, T.-R. and White, K. P. (2003). Tissue-specific gene expression and ecdysone-regulated genomic networks in drosophila. *Developmental cell*, 5(1):59–72.

[Li et al., 2008] Li, X.-y., MacArthur, S., Bourgon, R., Nix, D., Pollard, D. A., Iyer, V. N., Hechmer, A., Simirenko, L., Stapleton, M., Hendriks, C. L. L., et al. (2008). Transcription factors bind thousands of active and inactive regions in the drosophila blastoderm. *PLoS biology*, 6(2):e27.

[Love et al., 2014] Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550.

[Lubliner et al., 2015] Lubliner, S., Regev, I., Lotan-Pompan, M., Edelheit, S., Weinberger, A., and Segal, E. (2015). Core promoter sequence in yeast is a major determinant of expression level. *Genome research*, 25(7):1008–1017.

[Madrigal, 2015] Madrigal, P. (2015). On accounting for sequence-specific bias in genome-wide chromatin accessibility experiments: recent advances and contradictions. *Frontiers in bioengineering and biotechnology*, 3:144.

[McKay and Lieb, 2013] McKay, D. J. and Lieb, J. D. (2013). A common set of dna regulatory elements shapes drosophila appendages. *Developmental cell*, 27(3):306–318.

[McLean et al., 2010] McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M., and Bejerano, G. (2010). Great improves functional interpretation of cis-regulatory regions. *Nature biotechnology*, 28(5):495.

[McLeay and Bailey, 2010] McLeay, R. C. and Bailey, T. L. (2010). Motif enrichment analysis: a unified framework and an evaluation on chip data. *BMC bioinformatics*, 11(1):165.

[McLeay et al., 2012] McLeay, R. C., Lesluyes, T., Cuellar Partida, G., and Bailey, T. L. (2012). Genome-wide in silico prediction of gene expression. *Bioinformatics*, 28(21):2789–2796.

[Mugat et al., 2000] Mugat, B., Brodu, V., Kejzlarova-Lepesant, J., Antoniewski, C., Bayer, C. A., Fristrom, J. W., and Lepesant, J.-A. (2000). Dynamic expression of broad-complex isoforms mediates temporal control of an ecdysteroid target gene at the onset of drosophila metamorphosis. *Developmental biology*, 227(1):104–117.

[Natarajan et al., 2012] Natarajan, A., Yardımcı, G. G., Sheffield, N. C., Crawford, G. E., and Ohler, U. (2012). Predicting cell-type–specific gene expression from regions of open chromatin. *Genome research*, 22(9):1711–1722.

[Nègre et al., 2011] Nègre, N., Brown, C. D., Ma, L., Bristow, C. A., Miller, S. W., Wagner, U., Kheradpour, P., Eaton, M. L., Loriaux, P., Sealfon, R., et al. (2011). A cis-regulatory map of the drosophila genome. *Nature*, 471(7339):527.

[Neph et al., 2012] Neph, S., Vierstra, J., Stergachis, A. B., Reynolds, A. P., Haugen, E., Vernot, B., Thurman, R. E., John, S., Sandstrom, R., Johnson, A. K., et al. (2012). An

expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, 489(7414):83.

[Nishida et al., 2008] Nishida, K., Frith, M. C., and Nakai, K. (2008). Pseudocounts for transcription factor binding sites. *Nucleic acids research*, 37(3):939–944.

[Noyes et al., 2008] Noyes, M. B., Meng, X., Wakabayashi, A., Sinha, S., Brodsky, M. H., and Wolfe, S. A. (2008). A systematic characterization of factors that regulate drosophila segmentation via a bacterial one-hybrid system. *Nucleic acids research*, 36(8):2547–2560.

[Ou and King-Jones, 2013] Ou, Q. and King-Jones, K. (2013). What goes up must come down: transcription factors have their say in making ecdysone pulses. In *Current topics in developmental biology*, volume 103, pages 35–71. Elsevier.

[Ouyang et al., 2009] Ouyang, Z., Zhou, Q., and Wong, W. H. (2009). Chip-seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proceedings of the National Academy of Sciences*, 106(51):21521–21526.

[Pfreundt et al., 2009] Pfreundt, U., James, D. P., Tweedie, S., Wilson, D., Teichmann, S. A., and Adryan, B. (2009). Flytf: improved annotation and enhanced functionality of the drosophila transcription factor database. *Nucleic acids research*, 38(suppl_1):D443–D447.

[Piper et al., 2015] Piper, J., Assi, S. A., Cauchy, P., Ladroue, C., Cockerill, P. N., Bonifer, C., and Ott, S. (2015). Wellington-bootstrap: differential dnase-seq footprinting identifies cell-type determining transcription factors. *BMC genomics*, 16(1):1000.

[Piper et al., 2013] Piper, J., Elze, M. C., Cauchy, P., Cockerill, P. N., Bonifer, C., and Ott, S. (2013). Wellington: a novel method for the accurate identification of digital genomic footprints from dnase-seq data. *Nucleic acids research*, 41(21):e201–e201.

[Pique-Regi et al., 2011] Pique-Regi, R., Degner, J. F., Pai, A. A., Gaffney, D. J., Gilad, Y., and Pritchard, J. K. (2011). Accurate inference of transcription factor binding from dna sequence and chromatin accessibility data. *Genome research*, 21(3):447–455.

[Quinlan and Hall, 2010] Quinlan, A. R. and Hall, I. M. (2010). Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

[R Core Team, 2018] R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Rabinovich et al., 2016] Rabinovich, D., Yaniv, S. P., Alyagor, I., and Schuldiner, O. (2016). Nitric oxide as a switching mechanism between axon degeneration and regrowth during developmental remodeling. *Cell*, 164(1-2):170–182.

[Raj and McVicker, 2014] Raj, A. and McVicker, G. (2014). The genome shows its sensitive side. *Nature methods*, 11(1):39.

[Reinking et al., 2005] Reinking, J., Lam, M. M., Pardee, K., Sampson, H. M., Liu, S., Yang, P., Williams, S., White, W., Lajoie, G., Edwards, A., et al. (2005). The drosophila nuclear receptor e75 contains heme and is gas responsive. *Cell*, 122(2):195–207.

[Ritter and Beckstead, 2010] Ritter, A. R. and Beckstead, R. B. (2010). Sox14 is required for transcriptional and developmental responses to 20-hydroxyecdysone at the onset of drosophila metamorphosis. *Developmental Dynamics*, 239(10):2685–2694.

[Robinow et al., 1993] Robinow, S., Talbot, W. S., Hogness, D. S., and Truman, J. W. (1993). Programmed cell death in the drosophila cns is ecdysone-regulated and coupled with a specific ecdysone receptor isoform. *Development*, 119(4):1251–1259.

[Roeder, 1996] Roeder, R. G. (1996). The role of general initiation factors in transcription by rna polymerase ii. *Trends in biochemical sciences*, 21(9):327–335.

[Roider et al., 2006] Roider, H. G., Kanhere, A., Manke, T., and Vingron, M. (2006). Predicting transcription factor affinities to dna from a biophysical model. *Bioinformatics*, 23(2):134–141.

[Roy et al., 2010] Roy, S., Ernst, J., Kharchenko, P. V., Kheradpour, P., Negre, N., Eaton, M. L., Landolin, J. M., Bristow, C. A., Ma, L., Lin, M. F., et al. (2010). Identification of functional elements and regulatory circuits by drosophila modencode. *Science*, 330(6012):1787–1797.

[Sandelin et al., 2004] Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W., and Lenhard, B. (2004). Jaspar: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic acids research*, 32(suppl_1):D91–D94.

[Schmidt et al., 2016] Schmidt, F., Gasparoni, N., Gasparoni, G., Gianmoena, K., Cadenas, C., Polansky, J. K., Ebert, P., Nordström, K., Barann, M., Sinha, A., et al. (2016). Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. *Nucleic acids research*, 45(1):54–66.

[Schneider, 1972] Schneider, I. (1972). Cell lines derived from late embryonic stages of drosophila melanogaster. *Development*, 27(2):353–365.

[Schubiger et al., 2005] Schubiger, M., Carré, C., Antoniewski, C., and Truman, J. W. (2005). Ligand-dependent de-repression via ecr/usp acts as a gate to coordinate the differentiation of sensory neurons in the drosophila wing. *Development*, 132(23):5239–5248.

[Schweizer et al., 2003] Schweizer, L., Nellen, D., and Basler, K. (2003). Requirement for pangolin/dtcf in drosophila wingless signaling. *Proceedings of the National Academy of Sciences*, 100(10):5846–5851.

[Segal et al., 2008] Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U., and Gaul, U. (2008). Predicting expression patterns from regulatory sequence in drosophila segmentation. *Nature*, 451(7178):535.

[Segraves and Hogness, 1990] Segraves, W. A. and Hogness, D. S. (1990). The e75 ecdysone-inducible gene responsible for the 75b early puff in drosophila encodes two new members of the steroid receptor superfamily. *Genes & development*, 4(2):204–219.

[Shannon and Richards, 2017] Shannon, P. and Richards, M. (2017). Motifdb: An annotated collection of protein-dna binding sequence motifs. *R package*, 1(0).

[Shazman et al., 2013] Shazman, S., Lee, H., Socol, Y., Mann, R. S., and Honig, B. (2013). Onthefly: a database of drosophila melanogaster transcription factors and their binding sites. *Nucleic acids research*, 42(D1):D167–D171.

[Shlyueva et al., 2014] Shlyueva, D., Stelzer, C., Gerlach, D., Yáñez-Cuna, J. O., Rath, M., Boryń, Ł. M., Arnold, C. D., and Stark, A. (2014). Hormone-responsive enhancer-activity maps reveal predictive motifs, indirect repression, and targeting of closed chromatin. *Molecular cell*, 54(1):180–192.

[Sinha, 2006] Sinha, S. (2006). On counting position weight matrix matches in a sequence, with application to discriminative motif finding. *Bioinformatics*, 22(14):e454–e463.

[Sinha et al., 2003] Sinha, S., Van Nimwegen, E., and Siggia, E. D. (2003). A probabilistic method to detect regulatory modules. *Bioinformatics*, 19(suppl_1):i292–i301.

[Smale and Kadonaga, 2003] Smale, S. T. and Kadonaga, J. T. (2003). The rna polymerase ii core promoter. *Annual review of biochemistry*, 72(1):449–479.

[Song et al., 2011] Song, L., Zhang, Z., Grasfeder, L. L., Boyle, A. P., Giresi, P. G., Lee, B.-K., Sheffield, N. C., Gräf, S., Huss, M., Keefe, D., et al. (2011). Open chromatin defined by dnasei and faire identifies regulatory elements that shape cell-type identity. *Genome research*, 21(10):1757–1767.

[Spitz and Furlong, 2012] Spitz, F. and Furlong, E. E. (2012). Transcription factors: from enhancer binding to developmental control. *Nature reviews genetics*, 13(9):613.

[Stanojevic et al., 1991] Stanojevic, D., Small, S., and Levine, M. (1991). Regulation of a segmentation stripe by overlapping activators and repressors in the drosophila embryo. *Science*, 254(5036):1385–1387.

[Steinhauser et al., 2016] Steinhauser, S., Kurzawa, N., Eils, R., and Herrmann, C. (2016). A comprehensive comparison of tools for differential chip-seq analysis. *Briefings in bioinformatics*, 17(6):953–966.

[Stormo, 2000] Stormo, G. D. (2000). Dna binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23.

[Stormo et al., 1982] Stormo, G. D., Schneider, T. D., Gold, L., and Ehrenfeucht, A. (1982). Use of the 'perceptron' algorithm to distinguish translational initiation sites in e. coli. *Nucleic acids research*, 10(9):2997–3011.

[Sullivan and Thummel, 2003] Sullivan, A. A. and Thummel, C. S. (2003). Temporal profiles of nuclear receptor gene expression reveal coordinate transcriptional responses during drosophila development. *Molecular Endocrinology*, 17(11):2125–2137.

[Sung et al., 2016] Sung, M.-H., Baek, S., and Hager, G. L. (2016). Genome-wide footprinting: ready for prime time? *Nature methods*, 13(3):222.

[Sung et al., 2014] Sung, M.-H., Guertin, M. J., Baek, S., and Hager, G. L. (2014). Dnase footprint signatures are dictated by factor dynamics and dna sequence. *Molecular cell*, 56(2):275–285.

[Talbot et al., 1993] Talbot, W. S., Swyryd, E. A., and Hogness, D. S. (1993). Drosophila tissues with different metamorphic responses to ecdysone express different ecdysone receptor isoforms. *Cell*, 73(7):1323–1337.

[Tanay, 2006] Tanay, A. (2006). Extensive low-affinity transcriptional interactions in the yeast genome. *Genome research*, 16(8):962–972.

[Thomas et al., 2011] Thomas, S., Li, X.-Y., Sabo, P. J., Sandstrom, R., Thurman, R. E., Canfield, T. K., Giste, E., Fisher, W., Hammonds, A., Celniker, S. E., et al. (2011). Dynamic reprogramming of chromatin accessibility during drosophila embryo development. *Genome biology*, 12(5):R43.

[Thummel et al., 1990] Thummel, C. S., Burtis, K. C., and Hogness, D. S. (1990). Spatial and temporal patterns of e74 transcription during drosophila development. *Cell*, 61(1):101–111.

[Thurman et al., 2012] Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75.

[Truman et al., 1994] Truman, J. W., Talbot, W. S., Fahrbach, S. E., and Hogness, D. S. (1994). Ecdysone receptor expression in the cns correlates with stage-specific responses to ecdysteroids during drosophila and manduca development. *Development*, 120(1):219–234.

[Vierstra and Stamatoyannopoulos, 2016] Vierstra, J. and Stamatoyannopoulos, J. A. (2016). Genomic footprinting. *Nature methods*, 13(3):213.

[Vierstra et al., 2014] Vierstra, J., Wang, H., John, S., Sandstrom, R., and Stamatoyannopoulos, J. A. (2014). Coupling transcription factor occupancy to nucleosome architecture with dnase-flash. *Nature methods*, 11(1):66.

[Ward Jr, 1963] Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

[White et al., 1997] White, K. P., Hurban, P., Watanabe, T., and Hogness, D. S. (1997). Coordination of drosophila metamorphosis by two ecdysone-induced nuclear receptors. *Science*, 276(5309):114–117.

[Wu, 1980] Wu, C. (1980). The 5‘ ends of drosophila heat shock genes in chromatin are hypersensitive to dnase i. *Nature*, 286(5776):854.

[Wu et al., 2008] Wu, S., Liu, Y., Zheng, Y., Dong, J., and Pan, D. (2008). The tead/tef family protein scalloped mediates transcriptional output of the hippo growth-regulatory pathway. *Developmental cell*, 14(3):388–398.

[Yao et al., 2015] Yao, L., Berman, B. P., and Farnham, P. J. (2015). Demystifying the secret mission of enhancers: linking distal regulatory elements to target genes. *Critical reviews in biochemistry and molecular biology*, 50(6):550–573.

[Yu et al., 2015] Yu, G., Wang, L.-G., and He, Q.-Y. (2015). Chipseeker: an r/bioconductor package for chip peak annotation, comparison and visualization. *Bioinformatics*, 31(14):2382–2383.

[Zhang et al., 2008a] Zhang, L., Ren, F., Zhang, Q., Chen, Y., Wang, B., and Jiang, J. (2008a). The tead/tef family of transcription factor scalloped mediates hippo signaling in organ size control. *Developmental cell*, 14(3):377–387.

[Zhang et al., 2008b] Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., et al. (2008b). Model-based analysis of chip-seq (macs). *Genome biology*, 9(9):R137.

[Zhao et al., 2008] Zhao, B., Ye, X., Yu, J., Li, L., Li, W., Li, S., Yu, J., Lin, J. D., Wang, C.-Y., Chinnaiyan, A. M., et al. (2008). Tead mediates yap-dependent gene induction and growth control. *Genes & development*, 22(14):000–000.

[Zhu et al., 2010] Zhu, L. J., Christensen, R. G., Kazemian, M., Hull, C. J., Enuameh, M. S., Basciotta, M. D., Brasefield, J. A., Zhu, C., Asriyan, Y., Lapointe, D. S., et al. (2010). Flyfactorsurvey: a database of drosophila transcription factor binding specificities determined using the bacterial one-hybrid system. *Nucleic acids research*, 39(suppl_1):D111–D117.

[Zou and Hastie, 2005] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.