Transcriptional and epigenetic control of gene expression in embryo development

Ann Boija
To my family
Abstract

During cell specification, temporal and spatially restricted gene expression programs are set up, forming different cell types and ultimately a multicellular organism. In this thesis, we have studied the molecular mechanisms by which sequence specific transcription factors and coactivators regulate RNA polymerase II (Pol II) transcription to establish specific gene expression programs and what epigenetic patterns that follows. We found that the transcription factor Dorsal is responsible for establishing discrete epigenetic patterns in the presumptive mesoderm, neuroectoderm and dorsal ectoderm, during early Drosophila embryo development. In addition, these different chromatin states can be linked to distinct modes of Pol II regulation. Our results provide novel insights into how gene regulatory networks form an epigenetic landscape and how their coordinated actions specify cell identity.

CBP/p300 is a widely used co-activator and histone acetyltransferase (HAT) involved in transcriptional activation. We discovered that CBP occupies the genome preferentially together with Dorsal, and has a specific role during development in coordinating the dorsal-ventral axis of the Drosophila embryo. While CBP generally correlates with gene activation we also found CBP in H3K27me3 repressed chromatin.

Previous studies have shown that CBP has an important role at transcriptional enhancers. We provide evidence that the regulatory role of CBP does not stop at enhancers, but is extended to many genomic regions. CBP binds to insulators and regulates their activity by acetylating histones to prevent spreading of H3K27me3. We further discovered that CBP has a direct regulatory role at promoters. Using a highly potent CBP inhibitor in combination with ChIP and PRO-seq we found that CBP regulates promoter proximal pausing of Pol II. CBP promotes Pol II recruitment to promoters via a direct interaction with TFIIB, and promotes transcriptional elongation by acetylating the first nucleosome. CBP is regulating Pol II activity of nearly all expressed genes, however, either recruitment or release of Pol II is the rate-limiting step affected by CBP.

Taken together, these results reveal mechanistic insights into cell specification and transcriptional control during development.
List of Publications


Related publication


### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pol II</td>
<td>RNA polymerase II</td>
</tr>
<tr>
<td>TSS</td>
<td>Transcription start site</td>
</tr>
<tr>
<td>PIC</td>
<td>Pre-initiation complex</td>
</tr>
<tr>
<td>GTF</td>
<td>General transcription factor</td>
</tr>
<tr>
<td>TF</td>
<td>Transcription factor</td>
</tr>
<tr>
<td>CTD</td>
<td>C-terminal domain</td>
</tr>
<tr>
<td>CRM</td>
<td>Cis regulatory module</td>
</tr>
<tr>
<td>PTM</td>
<td>Post-translational modification</td>
</tr>
<tr>
<td>HAT</td>
<td>Histone acetyl transferase</td>
</tr>
<tr>
<td>HDAC</td>
<td>Histone deacetylase</td>
</tr>
<tr>
<td>H3K27ac</td>
<td>Histone 3 Lysine 27 acetylation</td>
</tr>
<tr>
<td>H3K27me3</td>
<td>Histone 3 Lysine 27 methylation</td>
</tr>
<tr>
<td>CBP</td>
<td>Creb binding protein</td>
</tr>
<tr>
<td>P300</td>
<td>300-kDa protein</td>
</tr>
<tr>
<td>GAF</td>
<td>GAGA-factor</td>
</tr>
<tr>
<td>RTS</td>
<td>Rubinstein-Taybi syndrome</td>
</tr>
<tr>
<td>ChIP</td>
<td>Chromatin immunoprecipitation</td>
</tr>
<tr>
<td>PRO-seq</td>
<td>Precision nuclear Run-On and sequencing</td>
</tr>
<tr>
<td>GRO-seq</td>
<td>Global Run-On and sequencing</td>
</tr>
<tr>
<td>STARR-seq</td>
<td>Self-transcribing active regulatory region sequencing</td>
</tr>
<tr>
<td>FAIRE-seq</td>
<td>Formaldehyde-Assisted Isolation of Regulatory Elements</td>
</tr>
<tr>
<td>PB</td>
<td>Pause button</td>
</tr>
<tr>
<td>DPE</td>
<td>Downstream promoter element</td>
</tr>
<tr>
<td>HMT</td>
<td>Histone methyl transferase</td>
</tr>
<tr>
<td>HDM</td>
<td>Histone demethylase</td>
</tr>
<tr>
<td>PcG</td>
<td>Polycomb Group proteins</td>
</tr>
<tr>
<td>TrxG</td>
<td>Trithorax Group proteins</td>
</tr>
<tr>
<td>PRE</td>
<td>Polycomb response element</td>
</tr>
<tr>
<td>MZT</td>
<td>Maternal-to-zygotic transition</td>
</tr>
<tr>
<td>ZGA</td>
<td>Zygotic genome activation</td>
</tr>
<tr>
<td>Sog</td>
<td>short gastrulation</td>
</tr>
<tr>
<td>Brk</td>
<td>brinker</td>
</tr>
</tbody>
</table>
# Table of Contents

Introduction .................................................................................................................. 11
Chromatin and transcription ................................................................. 12
  Transcription ........................................................................................................... 12
  Eukaryotic DNA is elegantly packed into chromatin ........................................... 13
  Transcriptional regulation ...................................................................................... 14

Key factors in Transcription ................................................................................. 15
  Pioneer factors open up the chromatin ................................................................. 15
  GTFs position Pol II at the promoter .................................................................... 16
  Co-regulators integrate regulatory signals during transcription ......................... 17

CBP/p300 coactivator .......................................................................................... 18
  Conservation .......................................................................................................... 18
  Domain structure of CBP/p300 ............................................................................. 18
  The role of CBP/p300 in disease .......................................................................... 19
  The role of CBP/p300 in regulating gene expression ........................................... 20

Key regulatory elements in Transcription .................................................... 22
  Promoter elements- determinants of the mechanisms of transcriptional control .......................................................................................................................... 22
  Enhancer elements- a key regulatory platform during transcription control .......... 24
  Insulator elements ................................................................................................. 27

Promoter proximal pausing of Pol II ......................................................... 28
  Pol II CTD .............................................................................................................. 28
  Pol II activity is regulated at multiple steps during transcription ....................... 29
  Key players of Pol II pausing ................................................................................ 30
  Experimental proof of Pol II pausing .................................................................... 30
  PRO-seq ................................................................................................................. 31
  The role of Pol II pausing ...................................................................................... 32

Histone modifications ..................................................................................... 33
  Chromatin states ................................................................................................. 34
  The role of histone modifications in transcription .............................................. 35
  Histone acetylation ............................................................................................... 36
  Histone acetyltransferases (HATs) ....................................................................... 36
  Histone methylation ............................................................................................. 37
  ChIP ......................................................................................................................... 38

Epigenetics ......................................................................................................... 39
  Establishment of cellular identity ....................................................................... 39
  Stability of cellular identity .................................................................................. 39
The epigenome is plastic................................................................. 40
Development – when an egg develops into an organism .......... 41
  Early embryo development ..................................................... 41
  Zygotic genome activation .................................................... 42
  Patterning of the embryo ..................................................... 42
  Anterior-posterior axis formation ........................................ 43
  Dorsal-ventral axis formation ................................................ 45
  Dpp signaling ......................................................................... 48
Aim of the thesis ........................................................................ 49
Results and discussion............................................................. 50
  Paper I .................................................................................. 50
  Paper II ................................................................................ 52
  Paper III ............................................................................... 53
  Paper IV .............................................................................. 54
Conclusion and perspectives .................................................. 55
Acknowledgements..................................................................... 56
Sammanfattning (Swedish summary): ...................................... 58
References ................................................................................. 60
Introduction

Cell specification is at the core of development, followed by cellular proliferation and maturation into a complete organism. Each cell in our body contains an identical set up of DNA. Despite this fact some cells will develop into muscle cells and others into neurons. This is a remarkable property of the genome, and the multifaceted processes forming regulated expression are both challenging and amazing to study. Distinct cells express a specific set of genes, i.e. a stretch of DNA that encodes for a function. Regulating which genes that are expressed result in the production of distinct set of proteins and thereby the unique properties of various tissues. These specific gene expression patterns need to be transferred to daughter cells during cell division to maintain their identical properties. In addition, during life, cells need to adapt to developmental and environmental cues and be able to change the expression status of particular genes. Understanding by what mechanisms gene expression programs are regulated is fundamental for cell identity and development but also for its misregulation in disease.
Chromatin and transcription

Transcription

Transcription is the process where the genetic information stored in DNA is transferred into RNA. The RNA is then spliced in the nucleus where introns are removed to produce messenger RNA (mRNA). The processed mRNA is transported to the cytoplasm and translated by the ribosome into protein. Adding post-translational modifications can in turn regulate the biological activity of the protein. This process of sequence information transfer is called the central dogma of molecular biology.

There are also noncoding parts of the DNA that are transcribed into transfer RNA (tRNA) and ribosomal RNA (rRNA) that are involved in the process of translating mRNA into proteins. Recent findings have identified numerous new non-coding RNAs that have been shown to play a significant regulatory role during transcription. Transcription is carried out by three types of RNA polymerases (Pol) (Roeder and Rutter, 1969). The different polymerases synthesize distinct sets of RNAs; Pol I (rRNA), Pol II (proteins, snRNA and miRNA) and Pol III (tRNA, rRNA and small RNA) (Weinmann and Roeder, 1974) (Weinmann et al., 1974). The three polymerases are very similar in subunit composition and structure, but the largest subunit Rpb1 of Pol II has an exclusive C-terminal domain (CTD) that plays a crucial role during transcriptional regulation.

The central dogma of transcription might seem like a linear and clear-cut process but has been shown to be remarkably complex. Transcription is a tightly regulated process critical for the state of virtually all cells, for cells to take on a specific identity during cellular differentiation and development as well as to avoid diseases.
Eukaryotic DNA is elegantly packed into chromatin

The eukaryotic cell must organize the DNA into a more compact form in order to fit the large amount of DNA into the tiny nucleus. This is achieved by wrapping 147bp of DNA in two super-helical turns around an octamer of histones (two of each H2A, H2B, H3 and H4) forming the basic unit of chromatin, the nucleosome (Kornberg, 1974) (Kornberg and Thomas, 1974) (Luger et al., 1997b). The compaction of DNA into chromatin forming the pattern of beads on a string is not only beautiful like a pearl necklace, it also serves as one important level of control during several essential cellular processes (Campos and Reinberg, 2009). Chromatin packaging limits the accessibility of the DNA and thereby influence key processes in the cell that utilizes the genetic information including gene regulation, DNA repair and replication. In addition, the elegant packaging of DNA into the well-structured form of chromatin mediates a highly dynamic structure. Rather than being a static form of packaging, histones arrange the DNA in a way that the sequence information can be stored but also used.

The link between chromatin and transcription comes from the discovery that nucleosomes hamper transcription in vitro (Workman and Roeder, 1987). It was later discovered that posttranslational modifications and histone remodeling including the addition or removal of histones as well as incorporation of histone variants could change the architecture of the chromatin (Li et al., 2007). This explosion in chromatin research provided new means of manipulating chromatin and thereby transcription, bringing the field in to the high-light.

**Genome organization.** DNA is wrapped around an octamer of histones, forming the basic package form, the nucleosomes. Nucleosomes are further organized into chromatin and finally into chromosomes.
Transcriptional regulation

Transcription factors (TFs) are DNA binding proteins that bind to regulatory elements of a specific set of genes to control their expression, either acting as activators or repressors. The targeting of a TF to its binding site is influenced by protein-protein interactions and the structure of the local chromatin. TFs interact with co-regulators that bridge interactions between TFs and the general machinery, composed of a group of general transcription factors (GTFs) and Pol II. Pol II is regulated at the step of recruitment to the promoter but also at the step of release from a paused state in order to go into productive elongation. Chromatin modifications recruit a specific set of chromatin regulators but also influence the accessibility of the chromatin to key DNA binding factors.

**Transcriptional regulation of a typical eukaryotic gene.** Transcription is regulated at several levels; binding of TFs and the recruitment of Coregulators, recruitment and release of Pol II and at the level of chromatin structure.
Key factors in Transcription

Pioneer factors open up the chromatin

Chromatin acts as a gatekeeper of regulatory regions by directing the accessibility of binding sites for transcription factors. Due to the dynamic nature of chromatin, the accessibility to DNA varies during development and in response to extracellular signaling. A special class of transcription factors, called pioneer factors, has the ability to bind nucleosomal DNA and open up the target region by recruiting chromatin-remodeling enzymes. Cooperative binding of pioneer factors and activators promote an open chromatin conformation and active transcription. By contrast, cooperative binding between pioneer factors and repressors result in closed chromatin and gene repression (Zaret and Mango, 2016).

**Action of pioneer factors.** Pioneer factors scan nucleosomal DNA. Upon binding of pioneer factors, they mediate the recruitment of other factors, either an activator or repressor resulting in gene activation or repression, respectively.
GTFs position Pol II at the promoter

Early *in vitro* studies revealed that isolated purified RNA polymerase together with nucleotides could synthesize RNA, but these RNAs were synthesized from random positions. Position specific RNA synthesis was achieved by adding crude nuclear extracts to the reaction (Luse and Roeder, 1980) (Weil et al., 1979). GTFs were later identified in human cells by separating proteins in whole cell extracts based on charge (Matsui et al., 1980). It revealed that none of the fractions alone, but the combination of fractions together could drive efficient transcription. The individual transcription factors were then discovered in the different fractions, Transcription Factor for RNA Pol II A (TFIIA) was located in fraction A, TFIID in fraction D and several factors in fraction C. The combined activity as well as the individual roles of the GTFs has been extensively characterized using biochemical studies and crystal structures (Woychik and Hampsey, 2002). The general transcription factors assemble at core promoter sequences (Thomas and Chiang, 2006) (Smale and Kadonaga, 2003) in a stepwise manner to initiate transcription (Orphanides et al., 1996). Promoter recognition is mediated by TATA-binding protein (TBP)-TFIID that recognizes core promoter elements and the interaction between TFIID and DNA is stabilized by TFIIA (Hoiby et al., 2007). TFIIB and TFIIF (Cabart et al., 2011) recruit Pol II and TFIIIE and TFIIH mediate promoter opening. The ATP-dependent translocase activity of TFIIH unwinds the DNA to form a transcription bubble. Holoenzyme complexes comprised of only a subset GTFs and Pol II have also been discovered (Chang and Jaehning, 1997) (Myer and Young, 1998). Regardless of the mechanism behind pre-initiation complex (PIC) assembly it will result in transcriptional initiation. However, the initiation may result in several rounds of short nascent transcripts that do not produce full-length transcripts.

**Formation of the pre-initiation complex (PIC).** Stepwise recruitment of GTFs and Pol II to form the PIC.
Co-regulators integrate regulatory signals during transcription

While the PIC contains all the minimal factors required for transcription it couldn’t respond to repressors, activators or transcribe genes in chromatin. For this purpose the PIC needs co-regulators and chromatin modifying complexes. Biochemical and functional studies have identified a large number of complexes that act as coregulators of specific transcriptional programs. Many coregulators are components of large multisubunit complexes with diverse enzymatic activity that can be summarized into two categories: histone modifying enzymes and ATP-dependent remodeling enzymes. Co-regulators associate with DNA binding factors and mediate its regulatory function. They can be divided into coactivators and corepressors with respect to their role during transcription.

Mediator
One widely used co-activator is the Mediator that acts as a bridge between regulatory factors and the PIC. However, the Mediator has also been shown to stimulate activator-independent transcription and has therefore been suggested to categorize as a GTF (Ansari et al., 2009) (Takagi and Kornberg, 2006). The Mediator has a flexible structure and can anatomically be divided into Head, Middle, Tail and Kinase domain. The Head is responsible for the contact with Pol II while the tail domain interacts with transcriptional regulators (Borggrefe and Yue, 2011). In contrast to GTFs, the Mediator has a global role in stimulating regulated transcription and exists in different subunit compositions (Malik and Roeder, 2010). Thus, in addition to its general activating function, the Mediator may regulate genes in a context specific manner.
CBP/p300 coactivator

The cyclic-AMP-response element binding (CREB) binding protein (CBP) and its paralog the 300-kDa protein (p300) are two widely used coactivators in metazoan cells. CBP (also called CREBBP and KAT3A) was originally discovered as a nuclear protein that interacted with the phosphorylated form of the transcription factor CREB to promote transcription (Chrivia et al., 1993). p300 (also called EP300 or KAT3B) was found when looking for interaction partners of the adenoviral oncogenic transcription factor E1A. CBP is composed of 2441 amino acids and p300 of 2412 amino acids and the two proteins share high sequence homology but not to other acetyltransferases, and are therefore collectively called CBP/p300 and comprise their own family of acetyltransferases.

Conservation

CBP/p300 have been identified in many different species belonging to higher eukaryotes. There are four CBP/p300 gene orthologs in Arabidopsis (Bordoli et al., 2001) and one copy in flies and roundworms. Rtt109 has been identified as a structural ortholog of CBP/p300 in yeast but lacks functional and sequence similarities (Tang et al., 2008). During the evolution of vertebrates, the chromosomal region of CBP/p300 has been duplicated resulting in the p300 gene at chromosome 22 and the CBP gene at chromosome 16 (Giles et al., 1998). The same composition of one CBP gene and one p300 gene can be seen in chicken, opossums, mice and humans, while frogs are missing the p300 gene (Dancy and Cole, 2015).

Domain structure of CBP/p300

CBP is a multidomain protein with a nuclear receptor interaction domain (NRID), several cysteine/histidine regions (CH), a CREB and MYB interaction domain (KIX), a bromodomain (binding acetylated lysines), a HAT domain (with an acidic surface that bind lysines and arginines and a nearby loop structure that binds the CoA substrate), a steroid receptor co-activator domain (SID) and an interferon response binding domain (IBiD) (Dancy and Cole, 2015). In addition, the N-terminal part of CBP has an ubiquitin ligase
function (Grossman et al., 2003) (Shi et al., 2009). The overall sequence similarity between CBP and p300 is 61%, however the acetyltransferase domain and its two flanking domains are especially enriched for sequence conservation (86%) (Chan and La Thangue, 2001). The many overlapping functions of CBP and p300 could be explained by the high sequence homology. For example, both proteins can bind to E1A and CREB but the two proteins also have distinct functions.

The overall sequence similarity between CBP and p300 is 61%, however the acetyltransferase domain and its two flanking domains are especially enriched for sequence conservation (86%) (Chan and La Thangue, 2001). The many overlapping functions of CBP and p300 could be explained by the high sequence homology. For example, both proteins can bind to E1A and CREB but the two proteins also have distinct functions.

Model of CBP/p300 domain structure. Nuclear receptor interacting domain (NRID), Three cysteine/histidine-rich (C/H) domains which also contain transcriptional adaptor zinc fingers (TAZ1) or a plant homeo domain (PHD), an interferon binding homology domain (IHD), a CREB and MYB interaction domain (KIX), a bromodomain, a HAT domain, a steroid receptor co-activator domain (SID), an interferon response binding domain (IBiD) and a proline containing PxP domain

The role of CBP/p300 in disease.

Altered expression of CBP/p300 results in numerous developmental defects. In humans these are manifested as a developmental disorder named Rubinstein-Taybi syndrome (RTS) (Petrij et al., 1995). Individuals with RTS exhibit short height, learning difficulties, broad thumbs and big toes and characteristic facial features including eyes and nose (Rubinstein and Taybi, 1963). The prevalence of the disease is 1 in 100 000 newborns and has been associated with alterations in the CBP gene (Hennekam, 2006). Human studies of CBP have been limited to cell culture models, however, an in vivo system is preferred when studying development. In flies, mice and worms, CBP/p300 is required for cell viability (Goodman and Smolik, 2000). Chromosome translocations that fuse monocytic leukemia zinc-finger protein (MOZ) with CBP or MLL-mixed lineage leukemia (MLL) with CBP have been associated with different types of leukemia. Somatic mutations in HAT- and C/H2-domain as well as truncated p300 are associated with loss of wild-type allele and tumor formation (Iyer et al., 2004). Frequent inactivating mutations in CBP and p300 have been found in several types of cancers, including small-cell lung cancer, bladder cancer, B cell lymphoma etc. (Shen and Laird, 2013).
The role of CBP/p300 in regulating gene expression

CBP is a widely used HAT and coactivator with more than 400 interaction partners (Bedford et al., 2010). CBP is a promiscuous HAT \textit{in vitro}, acetylating many proteins and histone lysines. However, CBP shows much greater substrate selectivity \textit{in vivo}. In cells, CBP and p300 are required to maintain global levels of H3K18ac and H3K27ac, but dispensable for other histone acetylations (Jin et al., 2011) (Tie et al., 2009). However, a recent study has shown that CBP is also responsible for acetylating H4K8 (Feller et al., 2015). In response to DNA damage, CBP mediates H3K56ac (Das et al., 2009). Furthermore, CBP is known to acetylate a total of 70 proteins \textit{in vivo} (Wang et al., 2008a).

CBP/p300 is well known for its role at enhancers, and genome wide mapping of CBP has been used to find novel enhancers both in human and flies (Visel et al., 2009) (Negre et al., 2011). Mapping CBP binding in different tissues can be used to predict tissue-specific activity of enhancers (Visel et al., 2009).

CBP has been suggested to regulate transcription by three main means, 1) CBP can act as an adaptor protein by bridging activators and GTF that subsequently mediate the recruitment of Pol II, 2) CBP can act as scaffolding protein facilitating interaction between proteins and DNA-and-proteins, 3) CBP can act as a HAT, acetylating both histones and non-histone proteins (Holmqvist and Mannervik, 2013).

The mechanism by which CBP/p300 is regulating transcription is the topic of paper IV.
The regulatory roles of CBP. CBP can regulate transcription by acting as (top) a bridge mediating the interaction between pre-initiation complex (PIC) and transcription factors (TFs), (middle) a scaffolding protein facilitating protein-protein interactions e.g. chromatin regulators (CRs) and DNA-protein interactions, (bottom), a HAT acetylating histones and non-histone proteins.
Key regulatory elements in Transcription

Promoter elements- determinants of the mechanisms of transcriptional control

The promoter is the central site of action to which the transcriptional machinery binds. Despite the fact that all promoters utilize PIC and that the factors are highly conserved, it is surprising that there is not a DNA element that is shared between all promoters. The core promoter is a region of about 100bp bordering the transcription start site (TSS) that is sufficient to drive correct transcriptional initiation and is composed of several core promoter elements that are recognized by the transcriptional machinery. The majority of promoters contain either of the two motifs TATA-box or Downstream promoter element (DPE) that are recognized by TFIID (Kutach and Kadonaga, 2000). The TATA-box is located -30 base pairs (bp) upstream of TSS and is an A/T rich region bound by TBP subunit of TFIID. DPE is found 30bp downstream of TSS and is bound by the TBP-associated factor (TAF) subunit of TFIID (Burke and Kadonaga, 1996). TFIIB recognition element (BRE) is situated just upstream of TATA-box and mediate an elevated affinity of TFIIB for the core promoter (Lagrange et al., 1998). Initiator (Inr) is situated right over the TSS and has a role in promoting correct transcriptional initiation (Smale and Baltimore, 1989). The motif 10 element (MTE) is positioned at +18 to +27 bp downstream of TSS and promotes transcription of Pol II (Lim et al., 2004). Pause bottom (PB) is located most commonly at +20 to +30 bp downstream of TSS (Hendrix et al., 2008) and is found at paused genes. The core promoter is extended by the proximal promoter, which holds a set of sequence-specific DNA-binding factors that impact transcription in different ways. One example of a proximal promoter motif is the GAGA motif commonly found at -100bp to -80bp of TSS, bound by GAGA-factor (GAF) and found at paused genes.

A subset of enhancers has a preference for a specific type of core promoter. Studies in the early Drosophila embryo have shown that enhancers in the Antennapedia gene complex and the Bithorax complex have a preference for activating TATA-containing promoters, while the rhomboid enhancer lacks a preference between TATA-containing and TATA-less promoters (Ohtsuki et al., 1998). Furthermore, a subset of enhancers has been found to only activate genes from a DPE containing enhancer and not a TATA containing
enhancer (Butler and Kadonaga, 2001). This indicates that the composition of the core promoter does not only drive the initiation of transcription but could also regulate enhancer function. One reason for having enhancer specificity to a core promoter could be to enhance accurate association of a distant enhancer and an explicit promoter within a promoter cluster.

Regulatory elements of a typical gene. The core promoter contains a selection of core promoter motifs that mediate its specific regulatory activity. The promoter proximal region holds binding sites for sequence specific DNA binding factors, e.g. GAGA-factor and heat shock factor (HSF). Enhancer elements also exhibit binding sites for sequence specific DNA binding factors and can be located at far distance.

Different promoter motifs have been associated with genes that have different rate-limiting steps during the transcription cycle. Genes that are regulated at the step of recruitment of Pol II to the promoter are rich in TATA-box motif (Chen et al., 2013). By contrast, genes that are or will be regulated at the step of release of the polymerase from the pause site are rich in PB and GAGA-motif (Chen et al., 2013). The Levine lab made a promoter swapping experiment, replacing the highly paused snail promoter with a less.paused promoter and found that it affected the synchrony of gene activation to a more stochastic activation, which resulted in variability of mesoderm invagination (Lagha et al., 2013). Thus, the composition of the promoter affects Pol II pausing and mode of gene activation and thereby key developmental processes. Furthermore, it has been shown that promoter motifs bound by Dref and GAGA-factor can separate between ubiquitously expressed housekeeping genes and tissue specific expression of developmental genes, respectively (Zabidi et al., 2015).

Taken together, the structure of core promoters is thus yet another layer of transcriptional control that contributes to the complexity of organisms.
Enhancer elements- a key regulatory platform during transcription control

While the promoter is sufficient to assemble the RNA pol II machinery and drive low levels of expression of its adjacent gene, high levels of expression often require activity from additional regulatory elements. These elements are called cis- regulatory modules (CRMs) and are located more distant from TSS, sometimes as far as 1Mb away. CRMs are also called enhancers, due to their role in upregulating, or enhancing transcription of target genes. The term enhancer was first coined after the discovery that the SV40 DNA could ectopically drive the expression of rabbit β-globin gene irrespective of orientation (Banerji et al., 1981). Later studies documented endogenous sequences in the immunoglobulin heavy chain locus with similar functions (Neuberger, 1983) (Gillies et al., 1983) (Banerji et al., 1983).

Enhancer function

Enhancers contain multiple short DNA motifs that act as binding sites for different tissue-specific transcription factors. These factors will in turn recruit co-activators and co-repressors and the sum of the combined regulatory activity of all factors bound will determine the transcriptional activity of the enhancer. Looping between enhancer and promoter element has been proposed to be critical for enhancer activity and would provide one possible explanation of how enhancers can exert their regulatory activity from far distance (Amano et al., 2009) Enhancer activity has also been associated with DNase I hypersensitivity (Boyle et al., 2008) as well as specific histone modifications of adjacent histones (Heintzman et al., 2009; Heintzman et al., 2007). Active enhancers are dressed with H3K4me1 and H3K27ac while flanking nucleosomes of inactive poised enhancers possess H3K4me1 and H3K27me3 (Shlyueva et al., 2014). Enhancer-derived RNAs (eRNAs) have been isolated but their function is poorly understood (Core et al., 2012) (De Santa et al., 2010). One possibility is that they are involved in keeping the chromatin accessible.
Enhancers can contact promoters from far distance. Upon binding of a tissue-specific transcription factor, coactivators are recruited and the enhancer and promoter are brought into close proximity by looping. Flanking nucleosomes of active enhancers are dressed with H3K4me1 and H3K27ac.

Global prediction of enhancers
The locations of enhancers are hard to predict due to the fact that they can be situated at various distance from their target genes. Advances in DNA sequencing during the last 10 years have discovered putative enhancers on a global scale. Several strategies in combination with sequencing have been used to predict enhancers on a genome-wide scale (Shlyueva et al., 2014). The chromatin landscape and the binding of regulatory proteins including transcription factors and co-factors has been mapped by ChIP-chip and ChIP-seq. Chromatin accessibility has been mapped using DNase-seq (Boyle et al., 2008), MNase-seq (Yuan et al., 2005) and Formaldehyde-Assisted Isolation of Regulatory elements (FAIRE-seq) (Giresi et al., 2007). The binding of CBP/p300, and the presence of H3K4me1 and H3K27ac have mainly been used to identify active enhancers.

Functional test of enhancers
Although the above studies have identified a large number of putative enhancers based on correlation with activity, few have actually been functionally tested on a global scale. Predicted enhancers have been tested in embryos by putting the candidate DNA sequence in front of a core promoter followed by a reporter gene. In situ hybridization can be used to determine the activity of the enhancer by measuring the abundance and localization of the reporter transcript in developing embryos. The enhancer activity could also be measured on the protein level by using enzymatic activity (β-galactosidase), fluorescence (GFP) and antibodies as a read out. These studies require the generation of transgenic animals and are therefore not suitable for enhancer screening genome-wide, but recent efforts have mapped thou-
sands of sequences for activity in vivo (Tomancak et al., 2007) (Kvon et al., 2014).

Advances in functionally testing enhancers have used deep sequencing to be able to test several enhancers in parallel (Shlyueva et al., 2014). Plasmid-based systems have also been used involving placement of the candidate sequence upstream of a minimal promoter followed by a reporter gene containing a barcode (Mogno et al., 2013). Self-transcribing active regulatory region sequencing (STARR-seq) brought the functionally testing of enhancers to a genome wide scale (Arnold et al., 2013). Since enhancers can drive the expression of target genes irrespective of orientation, the Stark lab put the enhancer sequence into the reporter gene downstream of the promoter. By directly linking the enhancer activity with its own sequence, and not by barcodes, STARR-seq allows the testing of millions of candidate sequences with variable length to be screened in batches. Systematic analysis of functionally characterized enhancer sequences have provided improved understanding of how regulatory function is encoded in the DNA. The draw back with plasmid-based systems is that you are not assessing the enhancer activity within the genome of a developing animal and might miss important aspects of developmental gene regulation at the chromatin level (Shlyueva et al., 2014). Recently, new methods to manipulate the activity of enhancers have arisen including transcription activator-like effectors (TALEs) (Crocker and Stern, 2013) and clustered regularly interspaced short palindromic repeat (CRISPR)-Cas9 system (Gilbert et al., 2013). In both systems, transcription factors and cofactors can be targeted to the site of interest and their regulatory effect can be monitored. In addition, the two systems can be used to edit the genomic DNA sequence (Reyon et al., 2012) (Jinek et al., 2012) (Mali et al., 2013). These methods will likely give further insights to the functional roles of individual key players.
Insulator elements

Insulator elements are DNA sequence elements that shield genes from inappropriate regulatory action from the surrounding. Insulators that are located between a promoter and an enhancer can block the ability of the enhancer to activate the promoter (Geyer and Corces, 1992) (Kellum and Schedl, 1991). However, the enhancer is free to activate promoters located on its insulator free side. The blocking ability of insulators has been suggested to be involved in structuring chromatin into topological domains with distinct functions (West et al., 2002). The insulator can also act as barriers by restricting the spread of condensed chromatin that could shut down the expression of genes (Sun and Elgin, 1999).

Insulators are bound by a specific set of insulator proteins for e.g. Su(Hw) (Parkhurst et al., 1988), BEAF (Zhao et al., 1995) and CTCF (Bell et al., 1999) that have been shown to mediate blocking function.

**Insulator activity.** Insulator elements protect genes from inappropriate regulatory action from the surrounding, (top) by blocking enhancer activity on promoters located behind the insulator element or (low) by acting as a barrier to prevent the spread of condensed chromatin.
Promoter proximal pausing of Pol II

The recruitment of the polymerase to the promoter was long believed to be the major rate-limiting step in transcription. However, more recent studies have shown that the release of a promoter proximal paused polymerase is the rate-limiting step at many genes. The Lis laboratory has contributed with insights on how Pol II pausing occurs by extensive studies on heat shock (hsp) genes in *Drosophila* (Gilmour and Lis, 1986) (Rougvie and Lis, 1988) (Giardina et al., 1992) (Rasmussen and Lis, 1993). The current view of the transcription cycle is 1. Recruitment of the polymerase to the promoter, 2. Transcriptional initiation. 3. Pausing of the polymerase immediately downstream of TSS. 4. Release from pausing resulting in productive elongation. 5. Termination of Pol II. 6. Re-initiation of transcription.

Pol II CTD

Modifying the activity of the central figure Pol II itself is one important layer of transcriptional regulation and occurs at several checkpoints. Pol II is a multi-subunit enzyme with a C-terminal domain (CTD) on the largest subunit, Rpb1, of Pol II. The protruding CTD has multiple heptad repeats and is the key target for regulation of Pol II activity. The number of heptad repeats varies between organisms, budding yeast contain 26-29 repeats, *Drosophila* 45 repeats and humans 52 repeats (Chapman et al., 2008). Regardless of the length of the CTD, the shared consensus sequence of the heptad repeats is $Y_1S_2P_3T_4S_5P_6S_7$. The CTD is subjected to the action of modifying enzymes that mediate various yet reversible modifications, where phosphorylation of Serines are the best studied. The existence of different forms of the polymerase was discovered on SDS-PAGE gels, which revealed two different motilities of Rpb1 (Schwartz and Roeder, 1975). One corresponded to a hypophosphorylated form of Pol II, which is recruited to the promoter and is a part of the PIC (Lu et al., 1991). The other is a hyperphosphorylated form which is associated with elongation (Cadena and Dahmus, 1987) and needs to be dephosphorylated in order to be recruited to the promoter again for another round of transcription (Cho et al., 1999).
Pol II activity is regulated at multiple steps during transcription

During the transcription cycle, the CTD is extensively modified at distinct steps (Buratowski, 2003, 2009). The three serines of the heptad repeat can be phosphorylated (Ser2, Ser5 and Ser7) (Corden, 2007; Palancade and Bensaude, 2003) which regulate the different steps of the transcription cycle. During PIC formation, Mediator bridges activators and unphosphorylated Pol II resulting in a fully assembled PIC at the promoter, stimulation of the kinase Cdk7 of TFIIH that mediates Pol II CTD phosphorylation at serine 5 and 7 and subsequent release of the Pol II CTD from the Mediator (Sogaard and Svejstrup, 2007) (Murakami et al., 2015). This result in promoter clearance, i.e. Pol II is detached from the PIC at the core promoter and starts to initiate transcription. At many genes, Pol II S5P is accumulated 20-50bp downstream of the TSS where it is held in a paused state. Phosphorylation of Pol II CTD Ser2 (Pol II S2P) is required for release into productive elongation. In addition, the CTD has an important regulatory role in serving as a platform for other enzymes participating in RNA maturation including protein complexes that cap, splice and polyadenylate RNA (Bentley, 2014) (Buratowski, 2009) (Egloff et al., 2012).

Non-canonical amino-acids are most prevalently found at position number seven of the heptapeptide repeat, where substituting canonical serine 7 to lysine 7 is the most common (Chapman et al., 2008). Lysine 7 is subjected to acetylation by p300 and associated with promoter proximal polymerases at paused genes in mammalian cells. CTD acetylation was further shown to specifically regulate the expression of growth factor induced genes (Schroder et al., 2013). Lysine 7 can also be methylated and associated with the earliest transcription stages before or together with serine 5 and 7 phosphorylation (Dias et al., 2015).

Given that lysine acetylation was critical for only a specific set of genes raises the question: How essential is the CTD for transcription? Early studies in yeast have shown that deleting parts of CTD down to less than 10 repeats are cell lethal, while 10-12 repeats are sufficient for conditional viability (Nonet et al., 1987). Mutation of Serine 7-p shows little effect on the expression of protein coding genes, but has shown to be important for transcription of short non-coding genes (Napolitano et al., 2014). Neither yeast mutants with impaired Ser2p show global effects on gene expression, but selectively affect genes with roles in meiosis (Sabarianfar et al., 2011).

Taken together, the regulatory function of Pol II CTD is highly complex. Different modifications of the CTD seem to play diverse roles on different sets of genes. Considering all possible CTD modifications of one heptad repeat and then the number of repeats it will give rise to a wide range of variations of different phosphorylation, acetylation and methylation patterns.
Key players of Pol II pausing

Sequence specific transcription factors associate with the promoter and work together with DRB sensitivity-inducing factor (DSIF) and the Negative elongation factor (NELF) to establish a paused polymerase. Upon recruitment of positive elongation factor b (P-TEFb) Ser2 of Pol II CTD becomes phosphorylated as well as NELF (resulting in eviction) and DSIF (transformed to a positive elongation factor). Recently, the crystal structure of mammalian Pol II in its transcribing form was resolved and positioned DSIF over the clamp domain of Pol II (Bernecky et al., 2016). P-TEFb is composed of cyclin T1 and cyclin-dependent kinase 9 (CDK9). In order for P-TEFb to promote transcription via its phosphorylating activity, it has to be released from an inhibitory complex where it is reversibly associated with a small nuclear ribonucleoprotein (snRNP) complex, 7SK (Zhou et al., 2012).

Experimental proof of Pol II pausing

Evidence of a paused polymerase comes from several different experimental techniques including permanganate footprinting that detects hypersensitivity of single stranded thymidines to oxidation. An open transcription bubble could be identified at 20-50 bp downstream of TSS at many genes (Gilmour and Fan, 2009). Recent advances in genome wide studies discovered that this is not just a phenomenon of a few genes but also a widespread signature of 10-40% of genes in human cells and Drosophila (Guenther et al., 2007) (Muse et al., 2007) (Zeitlinger et al., 2007). Chromatin immunoprecipitation sequencing (ChIP-seq) studies detect enrichment of cross-linked polymerase at the pause position. Sequencing of short capped RNAs reveal the presence of transcripts mapping to the region near promoters (Nechaiev et al., 2010). Nuclear run on assays, including GRO-seq and PRO-seq, have shown that the paused polymerase is still transcriptionally competent (Core et al., 2008) (Kwak et al., 2013).
The transcription cycle. Pol II is recruited to the promoter followed by TFIIH mediated phosphorylation of Pol II CTD on Serine 5, resulting in transcriptional initiation. Immediately downstream of TSS, Pol II is paused by NELF and DSIF. Upon recruitment of PTEF-b, Pol II CTD Serine 2 is phosphorylated and Pol II is released from pausing and proceeds into productive elongation. Pol II reaches the end of the gene and is terminated and can then be re-initiated to resume another round of the transcription cycle.

PRO-seq

Precision run-on sequencing (PRO-seq) is a technique that was developed in the Lis laboratory to map the levels and orientation of actively transcribing Pol II on a genome wide scale (Kwak et al., 2013). Nuclei are isolated and depletion of ribonucleotide monomers causes Pol II to stop transcribing but is kept in a transcriptional competent state. The run-on is performed by adding single biotin-labeled ribonucleotides that are incorporated by transcriptionally engaged polymerases and is performed in the presence of Sarkosyl that prevents new initiation of Pol II. A single nucleotide resolution is achieved due to the addition of only one of the four biotin labeled ATP/CTP/GTP/UTP that causes stalling of Pol II and prevents it from further transcription. Streptavidin beads are used for purifying the nascent RNA followed by deep sequencing The difference between PRO-seq and global run-on sequencing (GRO-seq) is that PRO-seq uses biotin-labeled ribonucleotides and GRO-seq BrUTP which results in a longer run-on and thereby decreased resolution (Core et al., 2008). The advantage with the PRO-seq technique is that in contrast to ChIP it does not rely on crosslinking efficiency or antibody specificity and detects transcriptionally engaged Pol II irrespective of phosphorylation status of Pol II CTD.
The role of Pol II pausing

While the existence and importance of Pol II pausing are no longer debated, the role of pausing is still not fully understood. Initially the purpose of paused Pol II was thought to be rapid activation of genes. The reason for this was that the heat shock genes were one of the first genes to be identified to possess a paused Pol II (Rougvie and Lis, 1988). Under high temperature, the heat shock genes are rapidly turned on and the paused polymerase is often believed to prepare the heat shock genes for fast induction. However, recent studies have shown that rapid induction does not always involve paused Pol II (Lin et al., 2011) and that Pol II pausing occurs more frequently in components of signaling cascades rather than inducible target genes (Gilchrist et al., 2012). Furthermore, while Pol II pausing is associated with active genes, its correlation with gene activity is poor (Min et al., 2011; Nechaev et al., 2010). Together these studies indicate that Pol II pausing has a role in mediating the potential of active elongation. At most genes Pol II pausing fine tunes the expression levels in response to environmental cues rather than acting as an on/off switch.

The binding of Pol II at the promoter proximal pause site has been suggested to be a mechanism by which the promoter region is kept nucleosome free (Gilchrist et al., 2010). Furthermore, Pol II pausing has been proposed to create a permissive state and is established in a stage specific fashion during development. Ultimately, this permissive state would allow the binding of tissue-specific transcription factor and when the right combination of transcriptional regulators is binding result in gene activation. Moreover, Pol II pausing is often present at multiple presumptive tissues in early development and is believed to mediate a permissive state that allows the response to morphogens resulting in gene activation in a subset of cells (Gaertner et al., 2012).

Pausing could serve as an extra step of transcriptional regulation. Due to the long residence time of Pol II pausing there is time for several interactions with transcriptional regulator that could modify the transcriptional state. Furthermore, the presence ofPaused Pol II RNA has been suggested to further allow interactions with epigenetic modifiers and transcriptional regulators.
Histone modifications

Chromatin modifications influence several nuclear processes including transcription. Post-translational modifications (PTMs) of histones can occur both on the globular domain but most prominently on the protruding histone tails. Several different amino acid residues of histones can be modified including lysine (K), arginine (R), serine (S), threonine (T), tyrosine (Y), by small structurally distinct moieties that include acetyl, methyl, phosphate and ubiquitin groups. Modifications of the protruding histone tails have been suggested to affect the inter-nucleosomal interaction and thereby the overall chromatin structure (Luger et al., 1997a). The discovery of acetylated histones and its association with high transcription of genes was made over 50 years ago (Allfrey et al., 1964). Since then, Chromatin immunoprecipitation (ChIP) has been used to map a wide range of histone modifications genome wide in various organisms during different developmental stages (Wang et al., 2008b) (Liu et al., 2005) (Negre et al., 2011) (Roudier et al., 2011). Recent reports have markedly expanded the list of histone modifications, and their role of action as well as the identification of novel modifications is under intense studies.

Modifications of the four core histones. Histones are dressed with various modifications including acetylation (ac), methylation (me), phosphorylation (Ph) and ubiquitination (Ub) on several different amino acids.
Chromatin states

Simplified, chromatin can be divided into two states, an open/active euchromatic state that is associated with histone acetylation and active transcription, and a closed/inactive heterochromatic state associated with histone methylation and repressed transcription. Heterochromatin can further be divided into constitutive heterochromatin marked by H3K9me2/3, which is found at repetitive DNA elements, and facultative heterochromatin marked by H3K27me3 that silences genes in a cell-type specific manner. However, recent studies have divided chromatin into more specialized groups.

**Chromatin states.** Chromatin can be divided into an acetylated euchromatin that is associated with active transcription, and an inactive heterochromatin state that is methylated. Heterochromatin can further be divided into H3K9me3 constitutive heterochromatin and H3K27me3 facultative heterochromatin.

With an attempt to functionally annotate the genome, the Encyclopedia of DNA Elements consortium has generated numerous genome wide data sets containing maps for histone modifications, chromosomal proteins, transcription factors, transcripts, replication proteins and nucleosome properties (Consortium, 2012). The same approach has been applied for model organisms, the model organism Encyclopedia of DNA Elements (modENCODE) project has generated more than 700 genome wide data sets containing maps for different developmental stages and several *Drosophila* cell lines. Computational approaches are then used to search for recurrent combinations that are grouped together and defined as chromatin states. The combination of 18 chromatin marks has been used to identify 9 broad classes of chromatin states (c1-c9), which can be further subdivided into 30 states (d1-d30), which showed enrichment for specific functions and regulatory elements (mod et al., 2010).

Since chromatin not only consist of DNA and histone proteins but also other chromosomal proteins, whole genome maps were generated of 53 chromatin proteins using DamID and led to the distinction of five chromatin states (Filion et al., 2010). Three states represent heterochromatin: GREEN (enriched in HP1 and H3K9me), BLUE (rich in PeG proteins and H3K27me3) and BLACK (enriched in non-coding elements, Histone H1 and low tran-
scriptional activity). Two states representing euchromatin: RED (enriched for e.g. Brahma and GAF) and YELLOW (rich in Mrg15 and H3K36me3). Taken together, whole genome mapping of chromatin factors and histone modifications can be used to group together recurrent combinations forming different chromatin states. The number of states is arbitrary and dependent on how broad or fine-grained one would like the classification to be. Regardless of the number of states defined, it is a simplification of tremendous amounts of information that is useful to understanding how the genome is organized. However, identifying the different states seems to be easier than to understand what they mean in the language of biology.

The role of histone modifications in transcription

Two models have been suggested on how histone modifications affect transcription. One model proposes that the regulatory action lies in the change in charge of histones that result in altered chromatin structure (Zheng and Hayes, 2003). H4K16ac has been shown to regulate higher-order chromatin structure by inhibiting the compaction of chromatin into 30nM fibers (Shogren-Knaak et al., 2006). The second model builds upon the existence of a “histone code” (Jenuwein and Allis, 2001). The combination of diverse histone modification would form different codes. The code would provide synergistic or antagonistic affinities for regulatory proteins that can affect the structure of the chromatin and regulate gene expression. Since information stored in chromatin can also be inherited through cell division, the role of chromatin within the field of epigenetics is an area under intense investigation.

Different types of histone modifications are localized to a specific position of a gene, together forming a histone landscape with distinct patterns. The precise location of histone modifications is essential for its regulatory output. For example, mistargeting of Set2 that normally methylates H3K36 in the gene body of transcriptionally active genes (Landry et al., 2003) results in gene repression (Strahl et al., 2002). Studies in yeast have revealed a number of hallmarks describing the chromatin landscape of a typical eukaryotic gene (Li et al., 2007). Nucleosomes have a higher density in the gene body than at the promoter, implying that promoter-binding sites of sequence specific transcription factors are located in more accessible regions (Bernstein et al., 2004) (Lee et al., 2004) (Sekinger et al., 2005). Active promoters are dressed with H3K4me3 and H3 and H4 acetylation. H3K36me3 and H3K4me2 are present in the coding region, where H3K4me1 is accumulating at the 3’ end. In addition, active or poised metazoan enhancers are enriched for H3K4me1 and H3K27ac (Creyghton et al., 2010). Studies in ES cells have also provided examples of bivalent domains that posses both the repressive mark H3K27me3 and the active mark H3K4me3 (Bernstein et al., 2006). A pro-
posed role for bivalent domains is to silence developmental genes, but keep them poised for activation at a later stage.

**Histone acetylation**

Histone acetylation has been described as a hallmark of chromatin dressing transcriptionally active genes. How acetylation promotes transcription is not fully understood but has been suggested to involve the weaker interaction between nucleosome and DNA caused by the neutralization of histone tails by acetylation, making them less attractive to the negatively charged DNA. Acetylation of different lysines is often found to coincide, H3K9ac, H3K18ac, H3K27ac and H4ac are found at TSS. In addition H4ac is also found throughout the gene of active genes (Wang et al., 2008b). This fits with the view that histone modifications may act cooperatively and that the cumulative effect of the number acetylated lysines make the gene ready for transcription (Li et al., 2007). The relaxation of the chromatin structure by acetylation mediates increased accessibility and thereby binding of transcription factors to their target sites. Alternatively, specific patterns of histone modifications may have distinct functions by directing regulatory factors to chromatin and could provide a mechanism for coordinated gene regulation (Kurdistani et al., 2004). Whether acetylation or not is a cause or consequence of transcription is still not clear (Roth et al., 2001). The acetylation status of lysines is a highly dynamic process governed by the antagonistic battle between HATs and HDACs. How these widely used proteins regulate specific genes as well as genes on a global level are ongoing questions in the field.

**Histone acetyltransferases (HATs)**

Histone acetyltransferases (HATs) are enzymes that add the reversible acetylation of specific lysine residues on histones. By contrast histone deacetylases (HDACs) removes acetyl groups from lysines resulting in a closed chromatin conformation. The first protein reported to have HAT activity was the transcriptional adaptor Gcn5 (Brownell et al., 1996). Gcn5 is part of the multisubunit complex SAGA, which can acetylate many lysine residues in vitro (Grant et al., 1999). The Saga complex is highly conserved and promotes transcription via four submodules with individual functions: HAT module acetylates histones, DUB module deubiquitinates H2B which promotes phosphorylation of Pol II CTD (Wyce et al., 2007), SPT module promotes PIC assembly and TAF module is important for the structure of SAGA (Koutelou et al., 2010). SAGA is targeted to chromatin via binding of Sgf29 to H3K4me2/3 resulting in H3K9ac and H3K14ac (Bian et al., 2011).
Since then, several other HATs have been discovered including CBP/p300, MYST (monocyte leukemia zinc-finger protein (MOZ), Ybf2, Sas2, Tip60) family of transcription factors and nuclear receptor coactivators. HATs are also referred to as KATs (Lysine acetyltransferase) due to their ability to also acetylate non-histone proteins. They are large multidomain and multi-protein complexes that are recruited to DNA by a wide variety of sequence specific transcription factors.

In *Drososphila*, complete loss of CBP is cell lethal and prevents oogenesis, but the hypomorphic allele nej1 manifest embryonic patterning phenotypes that can be explained by decreased Dpp-signalling caused by reduced expression of *tolloid* (*tld*) (Akimaru et al., 1997) (Lilja et al., 2003) (Waltzer and Bienz, 1999). The role of CBP/p300 in pattern formation is addressed in paper II.

Many questions still remain about how these enzymes regulate transcription. How are these broadly used HATs recruited to specific target genes? By what mechanism do HATs stimulate transcription? What are the genomic functions of HATs? These questions are addressed in paper II, III and IV.

Histone methylation

Histone methylation is both associated with an active and a repressive transcriptional state depending on where the methylation mark is located. Histone methyl transferases (HMTs) mediate the addition of methyl groups from S-adenosylmethionine to lysine that can be both mono- di- and trimethylated. By contrast, histone demethylases (HDMs) are responsible for the removal of methyl marks. Methylation of H3K9 is mediated by suppressor of variegation Su(var)3-9 (but also other HMTs) resulting in recruitment of HP1 and heterochromatin formation (Schotta et al., 2002). Polycomb Group (PcG) and Trithorax Group (TrxG) proteins are two key regulatory enzymes that mediate histone methylation but with antagonistic effects. TrxG proteins mediate H3K4 methylation that promotes transcription through recruitment of HATs and nucleosome remodelling complexes. PcG proteins have a repressive effect on transcription by forming H3K27me3 heterochromatin. 18 and 17 genes, respectively, have been assigned to PcG and TrxG proteins, and they tend to act in large protein complexes (Ringrose and Paro, 2004). Enhancer of zeste E(z) mediate the histone methyl transferase activity of Polycomb repressive comlex 2 (PRC2) and specifically methylates H3K27me3 (Czermin et al., 2002). Polycomb (Pc) of Polycomb repressive comlex 1 (PRC1) recognize this mark via its chromodomain (Cao and Zhang, 2004). The Trx subunit of TrxG proteins mediates the methylation of H3K4. The two protein complexes have a conserved core of proteins between *Drosophila* and human, but also unique accessory proteins. For example, the three sequence specific DNA-binding
proteins GAF, Pipsquek and Zeste are unique to *Drosophila* (Ringrose and Paro, 2004). PcG and TrxG target genes hold cis-regulatory elements termed Polycomb response elements (PRE) (Chan et al., 1994). Although target genes are recognized in human, PREs have not been identified. The two regulatory complexes were first identified for their role in *Drosophila* body patterning by regulating the expression of homeotic (*Hox*) genes (Ringrose and Paro, 2004). PcG and TrxG are important for maintaining the expression of *Hox* genes throughout development and adult life, and are therefore important in cellular memory.

**ChIP**

Chromatin immunoprecipitation (ChIP) is a widely used method for identifying binding sites of chromatin proteins. It involves crosslinking of your favorite protein to its chromosomal target sites in living cells, chopping up the cross-linked chromatin into smaller pieces to be able to immunoprecipitate the protein with a specific antibody. The crosslinking is reversed and you will isolate DNA that was bound by the protein of interest. The DNA can be analyzed by qPCR, hybridization to an array or more common today, by sequencing.

There are several crucial steps in order to achieve a successful ChIP-seq. Crosslinking needs to be done in a concentration and for a time that captures the protein but that does not overcrosslink resulting in artifacts and false peaks. Sonication of your chromatin in a size of about 100-300bp is suitable for sequencing and will give you high resolution. The antibody needs to be highly specific and preferentially verified by western blot and if possible, knockdown of your protein followed by ChIP. It is ideal to use two independent antibodies that generate consistent results. A recent report has identified non-specific enrichment in ChIP that is not seen in pre-immune serum or input, which they call phantom peaks (Jain et al., 2015). Phantom peaks are enriched in open chromatin of active promoters of highly expressed genes. The authors speculate that phantom peaks could represent regions that are particularly sticky in a combination with low ChIP specificity. When making sense of your genomic data, this is something to have in mind.
Epigenetics

The term epigenetics means above genetics and was originally coined by C.H. Waddington. His definition of epigenetics was “the causal interactions between genes and their products, which bring the phenotype into being” (Waddington CH. The epigenotype. Endeavour. 1942;1:18–20). Today one common definition of epigenetics is the heritable changes in gene expression that are independent of DNA sequence alterations.

Establishment of cellular identity

All cells in the body contain an identical set up of DNA but the expression status of genes differs between cells. Transcription factors have a central role in dictating the expression of a defined set of genes for each cell type. During cell differentiation, pluripotent cells receive input from regulatory factors resulting in induction of genes coding for additional transcription factors (Holmberg and Perlmann, 2012). Crossregulation of transcription factors will result in feedforward induction of specific factors that will activate each other and continue to work for an explicit cell fate. Complex transcription factor networks will give rise to a diverse set of gene expression programs, resulting in different cell types. Distinct expression programs of a cell will be accompanied by specific patterns of histone modifications. It might seem peculiar at first that the number of protein-coding genes in humans is only double the amount as that of the Drosophila fly. However, a more complex gene regulatory network could perhaps explain the greater complexity of me versus my flies in the lab. Increasing the number of transcription factors involved in regulating target genes would also expand the amount of different gene expression programs, resulting in several different types of cells and increased complexity of an organism.

Stability of cellular identity

Cell fusion and transcription factor induced reprogramming experiments have pointed out the importance of continuously active transcription factors for the maintenance of cell identity, a concept suggested already 25 years ago (Blau and Baltimore, 1991). Fusion of somatic cells have shown that
unknown factors in one cell can reprogram the fusion partner’s genome and induce cell-type specific gene activation (Blau et al., 1983). Forced expression of transcription factors can cause transdifferentiation, i.e. reprogramming that result in a switch of cell identity. Misexpression of myoblast determination protein 1 (MYOD1) in fibroblasts resulted in a conversion to skeletal muscle cells (Davis et al., 1987). Differentiated states are not always dependent on continuous instructions from key transcription factors. Instead, differential identities may be stabilized by DNA and chromatin modifications. One such example is the body segment identity in *Drosophila* that is controlled by *homeobox (hox)* and *engrailed* genes. Expression of key transcription factors in the early embryo set up the *hox* gene expression patterns, but TrxG and PcG proteins are responsible for maintaining them (Ringrose and Paro, 2004). Interestingly, PcG proteins have been implicated in regulation of several genes in the hierarchy that set up *hox* gene expression (McKeon et al., 1994) (Pelegri and Lehmann, 1994). This implies that transcription factor cascades may be supported by chromatin regulatory mechanisms and that this begins at an early point of the cascade.

The tight link between chromatin structure and regulation of gene expression is illustrated by Position-effect variegation (PEV). This was a phenomena discovered by H.J. Muller that upon the usage of X rays as a mutagen found variegated *Drosophila* eyes with areas of red and areas of white color (Henikoff, 1990). This phenotype was explained by chromosomal rearrangement that positioned the *white* gene, normally located in euchromatin, in close proximity to heterochromatin thus resulting in gene silencing. PEV has been reproduced for many genes when rearranged juxtaposed to heterochromatin (Girton and Johansen, 2008).

The epigenome is plastic

Numerous studies with diverse strategies have shown that a differentiated state is not irreversible, and that somatic differentiated cells can be reprogrammed all the way back to a pluripotent state. The first evidence comes from John Gurdon, who transferred a somatic differentiated cell nucleus into an enucleated *Xenopus* frog oocyte. He found that the somatic nucleus could be reprogrammed into pluripotency and had the capacity to develop into a complete animal (Gurdon, 1962). Somatic cells can also be reverted to a pluripotent state by expression of the key transcription factors Oct3/4, Sox2, c-Myc, and Klf4 (Takahashi and Yamanaka, 2006).
Development – when an egg develops into an organism

The development of no other animal with comparable or superior complexity is more comprehended than that of Drosophila melanogaster. In particular, a lot is known about the genetics of development and homology to Drosophila genes has been used to identify several key developmental genes also in other organisms. Using Drosophila as a model system has several practical advantages including, inexpensive to culture, short generation time of 10 days, use little space, and each female produce hundreds of eggs. In contrast to mouse and human where the embryo develops within the uterus, Drosophila embryo development is visible to study. The genome of Drosophila, which is condensed into 4 chromosomes, is exceptionally well annotated (Hoskins et al., 2015). In fact, 75% of disease causing genes in humans has counterparts also in fly, which makes findings in Drosophila also applicable to human biology and medicine (Reiter et al., 2001) (Chien et al., 2002) (Bier, 2005). The combination of genetic tools and the ChIP technology make the Drosophila embryo a good system to study transcriptional and epigenetic regulation during cell specification in vivo.

Early embryo development

During oogenesis, the egg is loaded with maternal mRNA and proteins, which direct the initial development of the early embryo. Upon fertilization, external stimuli activate the egg resulting in a rise of intracellular calcium that triggers the completion of meiosis, remodeling of the male pronucleus, reorganization of the cytoskeleton as well as an alteration in gene regulation. This results in fusion of the male and female pronucleus and is followed by rapid and synchronous cell cycles without gap phases. In Drosophila, the initial cell cycles occur without cytokinesis resulting in a syncytium with 6000 nuclei after 12 divisions, all sharing the same cytoplasm. This property of the syncytium mediates the possibility for key regulatory proteins to diffuse across nuclei during early Drosophila development that is of great importance during pattern formation. Following nuclear cycle 9, the nuclei migrate to the periphery where they undergo further syncytial divisions to form the syncytial blastoderm (represented by blastula or blastoderm stage in
other animals). After blastoderm formation, the cell cycles lengthen with inclusion of gap phases in a process termed mid-blastula-transition (MBT). During the Drosophila MBT the peripheral nuclei are enclosed by the plasma membrane to form a cellular blastoderm. About 15 nuclei located to the posterior end do not give rise to the cellular blastoderm but develop into pole cells and later germ cells that will form the sperm and egg.

**Zygotic genome activation**

During this initial part of embryogenesis, the mother provides all transcripts and proteins, but as development proceeds, the control is handed over to the zygote that will synthesis all required gene product from its own genome. The zygotic genome will be activated after a species-specific number of mitotic cycles, which is a conserved process in all animals, called the maternal-to-zygotic transition (MZT). The MZT can be divided into two distinct chapters, degradation of maternal products and activation of zygotic transcription (Tadros and Lipshitz, 2009). In Drosophila, the first part of zygotic genome activation (ZGA) occurs during cycle 8 and involves the activation of only a few tens of genes. The second part occurs at cycle 14 and engages several hundreds of zygotic genes into transcriptional activation.

Distinct groups of chromatin marks have been associated with the two different parts of ZGA. Acetylation of H4K8, H3K18, and H3K27 accumulate during the first part of ZGA while the chromatin is not dressed with K3K9ac, H3K4me1, H3K4me3, H3K36me3 and H3K27me3 until the second wave of transcription (Chen et al., 2013; Li et al., 2014). This temporal distinction of chromatin marks agrees with analysis of bulk levels of histone modification during Drosophila embryogenesis. High levels of H3K27ac are present during early stages of development while H3K27me3 is not detected until 4 hours of development (Tie et al., 2009). All together, these transcriptional program events result in dramatic morphological changes.

**Patterning of the embryo**

The importance of maternally provided RNA and proteins are well illustrated during patterning of the embryo. The principle behind animal embryo pattern formation is that initially equivalent cells in a developing tissue assume specific functions by responding to their position along a morphogen gradient. The first pattern is then further refined by short distance regulation resulting in fields of cells with the same properties. The morphogen gradient in the early embryo is composed of maternal gene products that specify the
body plan along the anterior-posterior and dorsal-ventral axes. Multifaceted regulations of the zygotic genes are used to further specify the embryo bodyplan. The two axes are organized more or less simultaneously but by different regulatory factors.

Anterior-posterior axis formation

During the onset of anterior-posterior axis formation, three maternal transcription factors Bicoid (Bcd), Hunchback (Hb) and Caudal (Cad) activate its target genes giant (gt), krüppel (kr) and knirps (kni) in spatial domains. Lack of these target genes causes gaps in the body pattern and is therefore collectively called gap genes. The anterior and posterior pole of the embryo is controlled by the activation of the receptor tyrosine kinase Torso (Li, 2005). Torso utilizes Ras-MAP signalling pathway to regulate the expression of the two repressors, huckebein (hkb) and tailless (tll). The expression pattern of gap genes is further refined by cross-regulation between Tailless, Giant, Krüppel and Knirps. Bicoid, Hunchback and Caudal also activate pair-rule genes whose expression is precisely limited to seven defined stripes perpendicular to the anterior-posterior axis by the repressive mechanism of gap genes. This tight regulation is mediated by two characteristics, 1. The CRM of pair-rule genes have high complexity with multiple binding sites for many regulatory factors 2. The regulatory function of Gap genes act only on a short range allowing individual CRMs to independently control transcription of the same target genes at the same time. The pair-rule genes in turn encodes for transcription factors that in a combinatorial fashion control the expression of the segment-polarity genes. The 14-stripe expression pattern of segment polarity genes is the precursors of body segments that finally acquire their unique character.
Patterning of the anterior-posterior axis in early *Drosophila* embryos. A hierarchy of factors set up the anterior-posterior axis. Maternally provided Bicoid (Bcd), Hunchback (Hb) and Caudal (Cad) activate gap genes in spatial domains. Bcd, Hb and Cad also activate pair-rule genes whose expression is restricted to seven stripes by the gap genes. The pair-rule genes in turn, regulate the expression of the segment-polarity genes into 14 stripes.
Dorsal-ventral axis formation

Patterning of the dorsal-ventral axis is ultimately controlled by the Rel/NFκB-family transcription factor Dorsal (Hong et al., 2008). The mother deposits Dorsal in the early embryo and is initially cytoplasmic but is translocated to the nucleus in response to Toll signaling. Pipe is synthesized on the ventral side of the embryo, by cells surrounding the oocyte (follicle cells). Pipe initiates an extracellular proteolytic cascade involving the proteases Nudel (Ndl), Gastrulation-defective (Gd), Snake (Snk) and Easter (Ea) resulting in cleaved and active Spätzle (Spz) ligand that binds to Toll receptor on the ventral side of the embryo. Toll activates Pelle kinase that in turn phosphorylates Cactus resulting in release of Dorsal protein that migrates into the nucleus and regulates its target genes. Taken together, this result in a gradient of Dorsal with highest concentration on the ventral side.

Dorsal regulates about 50 genes along dorsal-ventral axis in order to generate specialized tissues (Levine and Davidson, 2005). Dorsal activates genes in a concentration dependent manner, forming discrete thresholds of gene activity (Chopra and Levine, 2009). High levels of Dorsal at the most ventral side, activates snail (sna) and twist (twi) that possess enhancers with low-affinity Dorsal sites. Expression of these genes mediates the formation of the mesoderm that later becomes muscles and connective tissues. Intermediate to low levels of Dorsal activates short gastrulation (sog) and brinker (brk) that contains more optimal Dorsal sites as well as binding sites for the ubiquitously expressed transcription factor Zelda (Zld). Expression of these genes result in formation of the presumptive neuroectoderm and later the ventral nerve chord and ventral epidermis. Other neuroectoderm genes like ventral nervous system defective (vnd), rhomboid (rho) and vein (vn) have enhancers with fixed arrangement of Dorsal and Twist sites, mediating the response to intermediate levels of Dorsal. The Snail repressor prevents the expression of neuroectoderm genes from the ventral most cells. Absence of Dorsal activates decapentaplegic (dpp), tolloid (tll) and zerklüft (zen) (which are repressed by Dorsal in the mesoderm and neuroectoderm) and results in dorsal ectoderm formation. The repressive function of Dorsal involves the interaction with a protein that binds to AT-rich sequences flanking the Dorsal site. This result in a conformational change of Dorsal that expose a cryptic peptide motif (Ratnaparkhi et al., 2006) mediating the interaction with the co-repressor Groucho (Dubnicoff et al., 1997) (Valentine et al., 1998).

The concentration dependent response to Dorsal at target genes has been illustrated by using different mutants in the Toll signaling pathway (Anderson et al., 1985). Dominant mutation in the Toll gene (Toll<sup>10B</sup>) results in a constitutive active form of the receptor (Schneider et al., 1991) and
thereby the expression of \textit{sna} and \textit{twi} and an embryo that entirely consist of mesoderm. \textit{Toll}^{rm9/10} mutants have a partially active receptor (Schneider et al., 1991) and ubiquitously express \textit{sog} and \textit{brk} leading to neuroectoderm formation in the entire embryo. In \textit{pipe} mutants and \textit{gd} mutants which lack Toll signaling and are thereby devoid of Dorsal in the nucleus (Moussian and Roth, 2005), expression of genes such as \textit{dpp} and \textit{zen} form dorsal ectoderm in the entire embryo.

Taken together, Dorsal controls diverse transcriptional outputs by the combined effort of different affinity binding sites, number of binding sites and the co-binding with other transcription factors to mediate the many transcriptional outputs (Jiang and Levine, 1993).

The molecular mechanisms by which tissue-specific gene expression patterns are set up and how they are maintained has been a long-standing question and is the topic of paper I.
Patterning of the dorsal-ventral axis in early *Drosophila* embryos. Maternally provided Dorsal is present in a gradient with highest concentration on the most ventral side. Dorsal activates genes in a concentration dependent manner that subdivides the embryo into mesoderm, neuroectoderm and dorsal ectoderm. The Dorsal gradient is established by ventral localized Pipe that triggers a proteolytic cascade involving Nudel, Gastrulation-defective (Gd), Snake and Easter resulting in cleaved and active Spätzle ligand that binds to Toll receptor. Mutants in the Toll signaling pathway that have a homogenous population of cells with respect to Dorsal concentration are depicted to the right. *Toll*^10B^ mutants have a high Dorsal concentration and will form the mesoderm, *Toll*^m9/10^ mutants have low levels of Dorsal and will form the neuroectoderm and *gd* mutants lack dorsal in the nucleus and will form the dorsal ectoderm in the entire embryo.
Dpp signaling

Upon cellularization, transcription factors can no longer diffuse between nuclei and cell signaling is used for communication between cells. Dpp is a homolog of bone morphogenetic protein-4 (BMP-4) that belongs to the TGF-β family of cytokines that are involved in dorsal-ventral patterning also in vertebrates. Dpp acts as a morphogen with peak levels in the most dorsal part of the Drosophila embryo and specifies the dorsal ectoderm and the amnioserosa. Later in development, Dpp is involved in setting up wing imaginal disc patterning.

At the onset of cellularization, dpp is expressed uniformly in cells that lack dorsal in the nucleus but shortly becomes restricted to the most dorsal parts of the embryo. Dpp is inhibited by the two BMP related proteins Sog and Twisted gastrulation (Tsg) that bind Dpp and inhibit it from interacting with its receptor. Sog, a homolog of the vertebrate Chordin, is degraded when bound to Dpp by the metalloproteinase Tld resulting in release of Dpp. Dpp interacts with Screw and the two proteins signal through Thickveins (Tkv) and Saxophone (Sax) respectively. This result in activation of Smad-signaling where the class I Smad protein Mothers against dpp (Mad) becomes phosphorylated by Tkv and thereby associated with the class II co-Smad Medea (Whitman, 1998). The two Smads are translocated to the nucleus where they bind and activate their target genes. Brk acts as a repressor of Dpp target genes both in the early embryo and in the wing disc (Campbell and Tomlinson, 1999) (Jazwinska et al., 1999a) (Minami et al., 1999). Brk utilizes the long distance co-repressor Groucho (Gro) and the short distance co-repressor CtBP as well as its repressive domain 3R for its repressive action on target genes (Upadhyai and Campbell, 2013). In the early embryo, Brk together with Sog specify the neuroectoderm and restricts Dpp to more dorsal regions (Jazwinska et al., 1999b).
Aim of the thesis

The main objective of the thesis is to understand transcriptional and epigenetic regulation of gene expression. More specifically, it aims to describe how epigenetic patterns are first established during cell specification. It also dissects the molecular mechanisms by which the widely used co-activator CBP regulates gene expression and its role in development. The following questions are addressed:

- How is different epigenetic landscapes formed in the dorsal ectoderm, neuroectoderm and mesoderm in response to the Dorsal morphogen? (Addressed in paper I).

- What is the role of CBP at Dorsal target genes during early embryo development? (Addressed in paper II).

- What is the genomic function of CBP? What function does CBP bindings have outside of promoters and enhancers? (Addressed in paper III).

- How is CBP regulating the different steps in the transcription cycle, and specifically what is the role of CBP in promoter proximal pausing? (Addressed in Paper IV).
Results and discussion

Paper I

In order for a cell to take on a specific fate, transcriptional regulation and epigenetic mechanisms need to be tightly coordinated. The molecular mechanism for the initiation of specific transcriptional programs has been a subject of considerable interest, including the role of sequence specific transcription factor, polymerase and histone modifications in this process. We have performed ChIP-qPCR of key factors in transcriptional control at two neuroectoderm specific genes sog and brk, in naïve embryos and mutant Drosophila embryos that homogenously take on a specific dorsal-ventral fate forming the dorsal ectoderm, neuroectoderm and mesoderm.

We find that the pioneer factor Zelda is responsible for establishing a poised polymerase at Dorsal target genes prior to Dorsal activation in naïve cells. Pioneer factors are known for their role in cell differentiation by their unique property in being able to bind condensed chromatin. They endow competency for gene activity or repression by opening the chromatin to transcription factors and histone modifying enzymes. Our results indicate that some pioneer factors could mediate the establishment of a poised polymerase during the set up of different gene expression patterns. Several outstanding questions remain to be elucidated: By what molecular mechanism does pioneer factors mediate poised Pol II recruitment? What is the regulatory function of the poised polymerase in setting up tissue specific expression patterns? How frequent is this property among pioneer factors and at what types of genes?

Given that the process of genome activation is a highly conserved process, it will be interesting to further investigate if this is a general mechanism in animal development.

We also found that the key transcription factor Dorsal is responsible for setting up diverse epigenetic states. Upon cell specification, Dorsal activates its target genes in the neuroectoderm resulting in recruitment of the co-activator CBP/p300 and a hyperacetylated chromatin. In the mesoderm, sog and brk are not expressed despite the presence of Dorsal in the nucleus. The activity of a transcription factor can be regulated in several ways including competition for binding sites and protein-protein interactions. We find that Dorsal is still bound to these genes but this does not lead to CBP/p300 recruitment or acetylation of the chromatin. We propose that the Snail repressor quenches Dorsal activity leading to a hypoacetylated chromatin. It will be interesting
to further dissect by what mechanism Snail is modulating Dorsal activity. Furthermore, we find that in the dorsal ectoderm, these genes are held inactive by a Polycomb mediated mechanism involving H3K27me3 chromatin. It seems likely that Polycomb is recruited to these genes by a dorsal ectoderm specific factor, but how and what factor remains to be elucidated. Interestingly, we find that these two modes of repression result in different RNA polymerase regulation. Whereas a poised polymerase is still bound to the promoter in hypoacetylated chromatin, the polymerase is displaced from H3K27me3-chromatin. Thus, this provides a direct link between different modes of Pol II regulation and chromatin states.

It is well appreciated that epigenetic patterns are important for the maintenance of tissue-specific gene expression and used as a cellular memory to transfer gene activity programs to daughter cells. However, it is less known how these epigenetic patterns are first established during early development. Our results demonstrate how gene regulatory networks for nuclear programming initiate diverse epigenetic states to orchestrate embryogenesis.
Paper II

During embryo development, genes must be switched on and off in a highly coordinated and specific manner. Here we investigate the role of the widely used coactivator and histone acetyltransferase CBP/p300 in this process by using ChIP-seq and ChIP-qPCR in wt and mutant embryos lacking Dorsal in the nucleus. Given that CBP interacts and facilitates gene activation by multiple transcription factors from all major families, one might anticipate that CBP should be involved in setting up most transcriptional programs during development. Instead, we find that CBP is specifically regulating gene networks controlling dorsal-ventral patterning of the embryo. Hundreds of interaction partners have been described for CBP but how it is directed to DNA of its target genes is not known. We find that among 40 previously mapped transcription factors, CBP show strongest co-occupancy and also direct interaction with Dorsal. CBP occupancy is selectively Dorsal dependent at sites where few factors are bound. In the absence of Dorsal, CBP binding correlates best with Medea, a Smad-protein involved in Dpp-signaling, which is the opposing force that drives patterning of the dorsal-ventral axis.

HATs are well known for their role in stimulating transcription, but their broad role in transcriptional regulation of gene expression is not fully understood. We find that changes in CBP binding in mutant embryos correlated with the reciprocal change in gene expression levels genome wide. However, we also find that CBP is associated with silent genes and that the presence of H3K27me3 does not prevent the binding of CBP but constrains histone acetylation. Given that acetylation and methylation of H3K27 is mutually exclusive, this raises the possibility that CBP’s HAT activity could be regulated by substrate availability.

Whole-genome mapping of CBP in combination with H3K4me1 has been used to successfully predict enhancers. However, we find that CBP binding to many well-known active enhancers are below our high-confidence cut off, although the binding to many of them are above genome background. While mapping of CBP in different cell-types and developmental stages will increase the number of identified enhancers, our data also suggest that in order to get a full picture of regulatory sequences, CBP mapping should be used together with other approaches like mapping of additional HATs.

Taken together, CBP is a general coactivator with specific roles in development by coordinating the dorsal-ventral rather than for example the anterior-posterior axis of the embryo. CBP is also found at hypoacetylated silent genes. Like many other HATs, CBP is a big multidomain protein with several interaction opportunities with multiple regulatory factors, and could perhaps act as a scaffolding protein to regulate gene expression.
CBP is a widely used co-activator with HAT activity but its full role in regulating transcription and chromatin organisation is not known. We used CBP ChIP-seq in Drosophila S2 cells to map CBP binding genome wide and compared it with modENCODE data profiles of different chromatin proteins. We find that the majority of CBP is binding to active promoters and to both inactive/active enhancers consistent with a regulatory role of CBP in transcription. In addition we find that the strongest CBP binding is found at Polycomb regions that contain Polycomb proteins and H3K27me3 marks. At these sites, the binding of CBP does not result in activation or acetylation, indicating that CBP is not antagonizing H3K27me3 as is seen on a global level. Interesting for future studies would be to investigate if the HAT activity of CBP is blocked or if CBP is acetylating chromatin proteins but not histones. We also found that CBP is binding to insulator elements and regulates its insulating activity by acetylation of histones to prevent spreading of H3K27me3. In summary, CBP is bound to many genomic regions with different function including promoters, enhancers, insulators and Polycomb response elements (PREs). A common theme of these regions is that they directly interact with other genomic regions. Interestingly, this indicates that CBP might regulate higher order chromatin structure and the formation of chromosomal domains.
Paper IV

Transcriptional activation requires both the recruitment of the polymerase to the promoter and the subsequent release of Pol II into productive elongation. We used CBP ChIP seq and a highly potent CBP inhibitor in combination with ChIP-qPCR and PRO-seq to map Pol II at single nucleotide resolution to look at the role of CBP in the transcription cycle. We find that CBP is directly interacting with TFIIB and thereby promotes recruitment of Pol II. CBP is also regulating the release of polymerase into productive elongation by acetylating histones. This means that CBP has dual roles during transcriptional activation, mediating both of the key steps during the transcription cycle, recruitment and elongation.

We further found that CBP is regulating transcription on virtually all expressed promoters but that the rate-limiting step differs. At most promoters, the rate-limiting step is the release of the polymerase from the promoter into elongation. However, we also identified a special class of promoters with high levels of CBP and GAF that have higher levels of paused polymerase than other promoters. These highly paused promoters are marked by a unique set of histone modifications and chromatin factors, are highly expressed through development and largely shared between cell types. At these promoters the rate-limiting step affected by CBP is the recruitment of Pol II to the promoter.

Taken together, these results uncover a novel role of CBP at promoters in addition to its well-known role at enhancers. At expressed promoters, CBP has a global role in directly controlling promoter-proximal Pol II to stimulate transcription. It will be interesting to further dissect the role of CBP at promoters by investigating if CBP is acetylating key transcriptional regulators and examine how acetylation influence their action.
Conclusion and perspectives

A few final thoughts on transcription and epigenetic control during development are as follows. The strategies to regulate transcription are both complex and numerous, mediating the amazing diversity of life. While our understanding of how genes are regulated has greatly advanced, it has become evident that chromatin and transcription are two tightly linked key determinants in this process. Upon fertilization, the specialized germ cell is undressed from chromatin modifications in order for the two pronuclei to come together, forming a totipotent cell. Pioneer factors mediate the transition from a naïve cell to a differentiated cell by opening up the chromatin so other tissue specific factors can bind and recruit cofactors in order to initiate gene expression programs. The commencement of cell specification is accompanied by distinct histone modifications, first by acetylation and then by methylation that together form chromatin states that maintain proper expression patterns of different tissues. We have shown that the key transcription factor Dorsal is responsible for the establishment of different epigenetic landscapes on target genes in the three developing tissues along the dorsal-ventral axis of the embryo. The next step would be to map these epigenetic landscapes on a genome-wide scale.

Given the wide range of players discussed in this thesis that are involved in regulating transcription, understanding the molecular mechanism behind transcriptional control is a big challenge. However, the arrival of high-resolution methods to map chromatin proteins and transcriptionally engaged Pol II on a genome wide scale offer great opportunities to monitor these processes. In addition, the emergence of small molecule inhibitors provides a powerful tool to manipulate the function of specific key factors in transcription and in combination with genome wide methods will give direct and novel mechanistic insights to their role in gene regulation. Using these methods we discovered an unappreciated role of CBP/p300 at promoters in regulating Pol II pausing. Along with increasing mechanistic insights to transcriptional regulation a growing understanding of transcriptional misregulation is also being recognized. Small molecule inhibitors targeting key components in transcriptional control have already shown promising results as potential therapeutic targets in cancer.
Acknowledgements

There are many people who contributed to this work that I would like to express my gratitude to:

First and foremost, I would like to thank my supervisor Mattias Mannervik that with great expertise has provided me with generous guidance and support, but also for challenging me and urged me to do better. It’s been a pleasure to work with you!

Per-Henrik for invaluable hands on advises of how to elegantly perform experiments and for being a fantastic co-worker! Olle for being a generous person, giving good advises on cloning and fly-genetics and providing endless stories about pretty much everything. Filip for teaching me in situ and make excellent RNA preps. Together the four of us shared many years in the lab with endless discussion about science and life and we had a great deal of fun while doing it! Roshan for good collaboration and nice discussions. I would also like to thank all other members in the lab that I had the pleasure of working with Min, Suresh, George, Amanda, Bhumica, Kati, Helin and Olga.

Thanks to Christos Samakovlis, Stefan Åström and Qi Dai for their generosity in sharing resources and giving valuable advise on my work and career during the years. Monika for taking care of flies and making the department a nice place to work at.

Special thanks to Vasilios, for teaching me confocal microscopy and for our discussions about life.

Thanks to members of neighboring labs for sharing reagents and contributing to a nice working atmosphere: Ingrid, Naghmeh, Benedetta, Stefanie, Naveena, Emad, Jiang, Chie, Ryo, Georgia, Akkis, Alex, Ana, Dan, Enen, Andreas, Liqun, Katarina, Satish, Fergal, Badrul, Kirsten, Shenqiu, Chun and Taraneh.

Thanks are due to Stefan Björklund and Karl Ekwall for introducing me to research during my master studies, and the generous guidance from Nils Elfving, Miriam Larsson and Annelie Strålfors.
I am grateful for the productive collaboration with Per Stenberg and his team Philge and Aman. Big thanks to Per for taking the time to teach me how to run scripts and for advice on statistics!

Many thanks to John Lis and Jay for nice collaboration and long discussions about the role of CBP in pausing!

Many thanks to my family, extended family and friends for endless encouragement during these years.

Finally, I want to thank my husband Henric for never-ending support and for sharing happy and difficult times. Thanks to my kids Oscar and Greta for bringing so much love, joy and cheerfulness into my life.
Sammanfattning (Swedish summary):


I denna avhandling har vi studerat de molekylära mekanismerna som används för att etablera och bibehålla specifika genuttrycksprogram under embryoutevecklingen. Många cellulära processer inklusive hur genuttryck kontrolleras är konserverat mellan bananflugan Drosophila melanogaster och människa. Vi har därför använt oss av de avancerade genetiska verktyg som finns i bananflugen för att följa hur det epigenetiska mönstrum förändras på gennivå från naiva celler till specificerade vävnader under embryouteveckling. Vi har identifierat Dorsal som den faktor som både direkt och indirekt sätter upp olika epigenetiska mönster i celler som kommer att bilda muskelceller (mesodermet), nervceller (neuroektodermet) och i celler som kommer att bilda hudceller (dorsala ektodermet). Dorsal rekryterar CBP som acetylrerar histoner och leder till aktivt genuttryck. Vidare har vi påvisat två mekanismer som tyftar gener i dessa vävnader, antingen genom Polycomb medierad metylering av histone 3 lysin 27 eller genom hypoacetylerade histoner.
etablerade av Snail. Vi kan dessutom länka de olika histone modifieringsmönstren med distinkta sätt att reglera Polymeraset. Våra resultat tillhandahåller nya insikter till hur genregulatoriska nätverk formar det epigenetiska landskapet och hur dessa koordinerade aktioner specificerar cellidentitet.

CBP/p300 har en viktig roll i att reglera uttrycket av gener och mutationer av CBP är associerade med olika typer av cancer. Tidigare studier har visat att CBP har en viktig roll vid regulatoriska element kallade enhancers. Vi tillhandahåller bevis för att CBPs regulatoriska roll inte stannar vid enhancers utan sträcker sig till många genomiska regioner. Genom att använda en mycket potent CBP-drog har vi identifierat en ny regulatorisk roll hos CBP i att kontrollera ett nyckelsteg vid genaktivering: Polymerase pausing. CBP stimulerar rekryteringen av Polymeraset, till ett annat regulatoriskt element med namnet promotor, via direkt interaktion med TFIIB. CBP släpper iväg det pausade Polymeraset från promotorn längs genen genom att acetylera första oktameren av histoner. CBP reglerar Polymerasets aktivitet på i princip alla gener, men antingen rekrytering eller frigörandet av Polymeraset är det hastighetsbegränsande steget som påverkas av CBP.

Sammantaget bidrar dessa resultat till ny mekanistisk förståelse för transkriptionell och epigenetisk kontroll av genuttryck. Att kartlägga de underliggande mekanismerna som reglerar genuttryck är essentiellt för att förstå hur cancerceller skiljer sig mot vanliga celler.
References

methyltransferase activity that marks chromosomal Polycomb sites. Cell 111, 185-196.


regulators: the pervasive effects of Pol II pausing on stimulus-responsive
Do human chromosomal bands 16p13 and 22q11-13 share ancestral origins?
Am J Hum Genet 63, 1240-1242.
specific transcription enhancer element is located in the major intron of a
rearranged immunoglobulin heavy chain gene. Cell 33, 717-728.
polymerase in eukaryotic cells with permanganate genomic footprinting.
Methods 48, 368-374.
Gilmour, D.S., and Lis, J.T. (1986). RNA polymerase II interacts with the
promoter region of the noninduced hsp70 gene in Drosophila melanogaster
cells. Mol Cell Biol 6, 3984-3989.
FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates
active regulatory elements from human chromatin. Genome Res 17, 877-
885.
61, 1-43.
Grant, P.A., Eberharter, A., John, S., Cook, R.G., Turner, B.M., and
Workman, J.L. (1999). Expanded lysine acetylation specificity of Gcn5 in
(2007). A chromatin landmark and transcription initiation at most promoters
Gurdon, J.B. (1962). The developmental capacity of nuclei taken from
intestinal epithelium cells of feeding tadpoles. J Embryol Exp Morphol 10,
622-640.
Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A.,
Histone modifications at human enhancers reflect global cell-type-specific
Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins,
Distinct and predictive chromatin signatures of transcriptional promoters and
Hendrix, D.A., Hong, J.W., Zeitlinger, J., Rokhsar, D.S., and Levine, M.S.
(2008). Promoter elements associated with RNA Pol II stalling in the


embryonic stem cells mediated by the super elongation complex (SEC).
Genes Dev 25, 1486-1498.


Upadhyai, P., and Campbell, G. (2013). Brinker possesses multiple mechanisms for repression because its primary co-repressor, Groucho, may be unavailable in some cell types. Development 140, 4256-4265.

of a multiprotein repression complex at the ventral silencer. Mol Cell Biol 18, 6584-6594.