Genome-wide Organization of the Basal Transcription Factor TFIID

Thesis

How to cite:

© 2021 Sergio García-Moreno Alcántara

https://creativecommons.org/licenses/by-nc-nd/4.0/

Link(s) to article on publisher’s website:
http://dx.doi.org/doi:10.21954/ou.ro.00013249
Genome-wide organization of the basal transcription factor TFIID

Sergio García-Moreno Alcántara, M.Sc.

A thesis submitted in partial fulfillment of the requirements of the Open University for the Degree of Doctor of Philosophy

School of Life, Health and Chemical Sciences

The Stowers Institute for Medical Research
Kansas City, Missouri USA
an Affiliated Research Center of the Open University, UK

17-02-2021
Abstract

TFIID is a multi-subunit complex containing TBP and 13 TBP-associated factors (TAFs). It plays a key role in the initiation of transcription by recognizing promoter-specific sequence elements, leading to the assembly of the pre-initiation complex (PIC) machinery and the recruitment of the RNA-polymerase II. In vitro, TFIID uses different subunits to interact with such elements and changes its structural configuration when binding DNA. While motif composition is very diverse at endogenous promoters, whether all TFIID subunits are present across promoters or whether its binding organization changes in vivo with consequences for transcription is not understood. Here, we used ChIP-nexus, a technique that detects in vivo binding at base-pair resolution, to investigate the DNA-bound composition and organization of all 14 TFIID subunits in the Drosophila melanogaster genome. We found that the in vivo footprints of TFIID can be divided into three structurally distinguishable sub-modules whose binding is remarkably homogeneous across different promoter types. In contrast to the TAFs, the footprints for TBP were distinct for TATA, DPR, and housekeeping promoter types and revealed their underlying sequence elements. At most promoter types, TBP is found upstream and downstream of the start site dependent on the transcriptional state, but the amount and exact profile at each position differs between promoter types. We propose that TBP is first loaded downstream of the core promoter and then transferred upstream to assemble the transcription machinery, and that the time spent at each location depends on the promoter type and transcriptional activity. Taken together, our results suggest different rate-limiting states of TFIID recruitment between promoter types. We discuss the implications of this model for the transcription properties of each promoter.
Dedication

To my mother, for her tireless spirit
Acknowledgements

First of all, I would like to deeply thank Julia Zeitlinger, for her advice and guidance, and for teaching me that creativity is a fundamental part of being a scientist. I cannot imagine being where I am without her support and mentorship. I am also truly grateful for all the people that I have met and have helped me along the way. A special thank you to Wanqing Shao, from who I learnt all the ChIP-seq and nexus experiments and the genomic data analysis. To Cindi Staber for all kinds of lab support and design of the TAFs antibodies. To Melanie Weilert, for all her help with the data analysis. To Anna Kogan and Kaelan Brennan for contributions in the ChIP-nexus experiments and embryo collections. To Sara Jackson for administrative assistance. To Nilay Sah, for being a nice buddy and to all the other Zeitlinger lab members for creating such a supportive environment.

I am also thankful to my Ph.D. advisory committee members Joan Conaway, Jerry Workman and Ariel Bazzini for their advice and support throughout this project. I also want to thank Leanne Wiedemann for making all the thesis-related bureaucracy much smoother and to the Open University for the program and for their support towards the completion of this thesis.

Last but not least, I really want to thank my mother and all my friends, especially to Ruben Hervás and Ricary Osiro, I am deeply grateful to have encountered them in my life and I could not imagine completing this journey without their support and friendship.
Table of Contents

Abstract ii
Dedication iii
Acknowledgements iv
Table of Contents v
Table of Figures vii
Table of Tables viii

Chapter 1 Introduction 1
1.1 DNA, Histones and Chromatin 2
1.2 The histone code and chromatin dynamics 6
1.3 RNA polymerase II transcription 8
1.4 Basal transcription factors are required for accurate Pol II initiation 12
1.5 Pol II pausing, a rate-limiting step for metazoan gene transcription 16
1.6 The core promoter region: at the heart of transcription 19
1.7 The TFIID complex is a specialized core promoter recognition machine 26
1.8 Composition and assembly of the TFIID complex 27
1.9 Assembly of the holo-TFIID complex 40
1.10 Structural basis of TFIID promoter recognition 43
1.11 Genome-wide mapping of transcriptional complexes at high-resolution can be used to determine their DNA-bound organization 48
1.12 Aims of this thesis 52

Chapter 2 Results 54
2.1 Holo-TFIID is present at all active promoters 54
2.2 TFIID organization at focused vs dispersed promoters 61
2.3 The binding profile of TBP, but not TAFs, is promoter-specific 64
2.4 TBP appears to bind downstream of the TSS before PIC assembly 68
2.5 Analysis of TBP and Pol II ChIP-nexus data at dual promoters during early Drosophila embryogenesis reveals no association between TBP and pausing 72
2.6 TFIID is released from the promoter during Pol II pausing 76

Chapter 3 Discussion 81
3.1 Genome-wide composition of the TFIID complex 81
3.2 Genome-wide spatial organization of the TFIID complex 83
3.3 TFIID binding during the transcription cycle and consequences for the transcription dynamics at different promoter types 87

Chapter 4 Conclusions & future directions 98

Chapter 5 Materials & Methods 101
5.1 Materials 101
5.2 Methods 105
   Cell culture 105
   ChIP-nexus experiments 105
   RNA-seq experiment 107
   Embryo collections 107
5.3 Quantification and statistical analysis 107
   TSS annotation of Drosophila melanogaster kc167 cells 107
   Occupancy heatmaps 108
   Correlation heatmap 108
   Average binding profiles 108
   Focused vs Dispersed promoters 109
   Selection of promoter types 109
   Total signal distributions (occupancy boxplots) 110
   Pol II pausing half-life values 111
   Changes in TFIID occupancy after TRI 111
   RNA expression levels 112
   Public data sets used for analysis 112
## Table of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1-1</td>
<td>DNA structure and transcription factor binding</td>
<td>3</td>
</tr>
<tr>
<td>Figure 1-2</td>
<td>Histone proteins and nucleosome assembly</td>
<td>5</td>
</tr>
<tr>
<td>Figure 1-3</td>
<td>Overview of common histone PTMs</td>
<td>7</td>
</tr>
<tr>
<td>Figure 1-4</td>
<td>Consequences of ATP-dependent chromatin remodelling</td>
<td>7</td>
</tr>
<tr>
<td>Figure 1-5</td>
<td>Pol II and the CTD phosphorylation cycle</td>
<td>11</td>
</tr>
<tr>
<td>Figure 1-6</td>
<td>Stepwise assembly and structure of the PIC</td>
<td>15</td>
</tr>
<tr>
<td>Figure 1-7</td>
<td>Schematic overview of the establishment of Pol II pausing and release</td>
<td>18</td>
</tr>
<tr>
<td>Figure 1-8</td>
<td>Core promoters differ in sequence composition, initiation pattern and function</td>
<td>24</td>
</tr>
<tr>
<td>Figure 1-9</td>
<td>Hypothetical model for TFIID assembly</td>
<td>43</td>
</tr>
<tr>
<td>Figure 1-10</td>
<td>The structure of TFIID and model for DNA binding</td>
<td>47</td>
</tr>
<tr>
<td>Figure 1-11</td>
<td>Common methods to detect TF location in vivo</td>
<td>51</td>
</tr>
<tr>
<td>Figure 2-1</td>
<td>DNA-bound TFIID is organized into three major modules</td>
<td>59</td>
</tr>
<tr>
<td>Figure 2-2</td>
<td>TFIID is more stable at focused/regulated promoters</td>
<td>63</td>
</tr>
<tr>
<td>Figure 2-3</td>
<td>The binding profile of TBP is promoter specific</td>
<td>66</td>
</tr>
<tr>
<td>Figure 2-4</td>
<td>Clustering genes based on TBP binding profile distinguishes between promoter types</td>
<td>67</td>
</tr>
<tr>
<td>Figure 2-5</td>
<td>TBP binds downstream of the TSS before PIC assembly</td>
<td>71</td>
</tr>
<tr>
<td>Figure 2-6</td>
<td>Downstream TBP does not associated with the paused Pol II</td>
<td>74</td>
</tr>
<tr>
<td>Figure 2-7</td>
<td>Downstream TBP is also present in the Drosophila embryo</td>
<td>75</td>
</tr>
<tr>
<td>Figure 2-8</td>
<td>TFIID occupancy correlates better with initiation than pausing</td>
<td>77</td>
</tr>
<tr>
<td>Figure 2-9</td>
<td>TFIID binding is stabilized during PIC assembly</td>
<td>80</td>
</tr>
<tr>
<td>Figure 3-1</td>
<td>Model for TFIID binding dynamics at different promoter types</td>
<td>97</td>
</tr>
</tbody>
</table>
### Table of Tables

Table 1-1. Core promoter elements summary ................................................................. 25  
Table 4-1. Cell lines and fly stocks ................................................................................. 101  
Table 4-2. Antibodies ........................................................................................................ 101  
Table 4-3. Standard ChIP-buffer ..................................................................................... 102  
Table 4-4. ChIP-nexus wash buffer A .............................................................................. 102  
Table 4-5. ChIP-nexus wash buffer B ............................................................................... 102  
Table 4-6. ChIP-nexus wash buffer C ............................................................................... 102  
Table 4-7. ChIP-nexus wash buffer D ............................................................................... 102  
Table 4-8. Tris buffers ....................................................................................................... 102  
Table 4-9. RIPA buffer .................................................................................................... 102  
Table 4-10. TE buffer ...................................................................................................... 102  
Table 4-11. Elution buffer ............................................................................................... 102  
Table 4-12. Chemicals and reagents ............................................................................... 103  
Table 4-13. ChIP-nexus oligos ....................................................................................... 104  
Table 4-14. Core promoter elements information ........................................................... 110
Chapter 1 Introduction

At any given time, all living organisms are exposed to a plethora of fluctuating environmental signals and intracellular cues. To survive, cells must respond to those signals and be able to adjust their metabolism and behavior to new circumstances. Most often, this involves modulating gene expression programs in different cells and tissues, leading to changes in the amount of RNA and proteins within the cell.

Proteins can be defined as “molecular machines”. They carry out the majority of cellular processes, including gene expression. Because proteins are synthesized from an mRNA template, which has been previously transcribed from a DNA sequence, the regulation of mRNA transcription becomes the primary mechanism controlling protein synthesis and is therefore crucial for the maintenance of cellular homeostasis.

Over half a century ago, Bob Roeder and colleagues discovered the RNA polymerase II and a set of basal transcription factors (TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH) as the key effectors required for mRNA synthesis. Since then, our understanding of eukaryotic transcription has expanded exponentially, revealing a chaotic world of tightly regulated processes. In this chapter, I will review the main steps required for accurate RNA polymerase II transcription and summarize the main factors involved. I will then focused on the role of the transcription factor II D (TFIID), the subject of this thesis work and who, upon binding to DNA, marks the beginning of one of the most important cellular processes – gene transcription.
The discovery of DNA as the container of genetic inheritance, together with the unveiling of its structure in the early 50s, were among the most laurate scientific discoveries of the last century. The resolution of the DNA structure was possible thanks to the landmark work of Rosalind Franklin, whose x-ray picture of DNA was essential for James Watson and Francis Crick to postulate their famous double-helix arrangement of the DNA molecule in 1953 (Cramer 2020).

Chemically, the DNA is a polymer of nucleotides composed of a deoxyribose-phosphate backbone and four different nitrogenous bases: Adenine (A), Timine (T), Guanine (G) and Cytosine (C). Within the double-helix, the DNA strands are positioned in an antiparallel manner and show complementary pairing between A-T and G-C nucleotides on each side (Franklin and Gosling 1953b; Franklin and Gosling 1953a; Watson and Crick 1953; Maddox 2003). This offset arrangement causes a slight distortion acting in direct opposition between the two helical strands, giving rise to the major and minor DNA grooves, which run opposite to each other along the entire DNA molecule (Wang 1979) (Figure 1-1A). These architectural features are fundamental for the accessibility to DNA by regulatory proteins and represent by themselves a layer of transcriptional regulation (Privalov et al. 2007) (Figure 1-1B).

In eukaryotes, all the genetic information is stored within the cell nucleus. However, the length of DNA fibers greatly overpasses the recipients in which they are contained. In humans, for example, if one were to stretch the entirety of the genetic material found in one single cell, would end up with a two-meter long fiber in their hands (Venter et al. 2001), meaning that the DNA molecule has to be extensively compressed in order to fit within the nucleus (Phillips and Milo 2009). This is attained by wrapping the double DNA helix around a group nuclear protein named nucleosomes, which at low resolution look like “beads on a
string” when imaged using electron microscopy (EM) (Kornberg 1974; Olins and Olins 1974; Oudet et al. 1975). Together, DNA and nucleosomes form the basal state in which the genetic information is found inside the cell – the chromatin.

![DNA structure and transcription factor binding](image)

**Figure 1-1. DNA structure and transcription factor binding**

(A) Fragment of a DNA double-helix showing its main structural features. (B) The CCAAT-binding complex (CBC), an important factor for gene expression, recognizes the CCAAT-box located in the minor groove. Adapted from (Anon n.d.; Huber et al. 2012)

For almost 100 years after their discovery in 1884, nucleosomes were merely considered as inert packing material for eukaryotic DNA. It was not until the decade of 1980, when the pioneer work by Roger Kornberg, Michael Grunstein and David Allis, revealed these nuclear proteins as active players of gene regulation. Itself, the nucleosome is an octamer of histones – basic proteins highly conserved through evolution (Luger et al. 1997; McGinty and Tan 2015). There are five types of histones in eukaryotes: H1, H2A, H2B, H3 and H4, but only the last four, termed the core histones, are part of the octamer, while H1 works as a linker between nucleosomes in higher levels of chromatin organization (Thoma et al. 1979; Hergeth and Schneider 2015; Robinson and Rhodes 2006).

All core histones are present in two copies and form heterodimers (H3-H4 and H2A-
H2B) within the nucleus (Luger et al. 1997; McGinty and Tan 2015). Their assembly into a stable octamer is driven by the strong affinity between histones due to the presence of a histone fold domain (HFD) within their structure (Figure 1-2A). This domain consists of three α-helices linked by two unstructured loops and connects the histone pairs in a head-to-tail fashion, a characteristic interaction that resembles a “molecular handshake”.

Another common feature among histones is a “helix-turn-helix” domain that allows interaction with DNA. It is worth mentioning that histones contact DNA primarily at its backbone and could, theoretically, assemble nucleosomes in a sequence-independent manner. However, different nucleotide combinations affect the physical properties of the DNA locally, making it more or less energetically favorable to assemble nucleosomes. For example, AA, TA, TT (and similar with GCs) dinucleotides with 10 bps spacing periodicity drive a nucleosome-favorable rotational positioning of DNA, while long stretches of As or Ts (poly A/T tails) are not prone to DNA bending and therefore disfavors binding of nucleosomes (Thåström et al. 1999; Anselmi et al. 1999).

When the conditions are favorable, nucleosome formation begins with the assembly of an H3/H4 tetramer and its association with DNA, followed by the incorporation of two separate H2A/H2B heterodimers and completion of the double helix wrapping (Wilhelm et al. 1978; Burgess and Zhang 2013) (Figure 1-2B). Importantly, several histone variants can replace their canonical counterparts and support the nucleosome structure (Henikoff and Smith 2015; Talbert and Henikoff 2017). These non-canonical histones alter nucleosome stability and are essential for the establishment of both, high-condense chromatin regions inaccessible to transcription factors, or loosely bound and accessible DNA regions. Two non-canonical histones important for transcription are the H3.3 and H2A.Z. They appear close to the transcription start site of genes and stimulate transcription by destabilizing histone-histone and DNA-histone interactions, allowing for the binding of transcription factors.
Figure 1-2. Histone proteins and nucleosome assembly

(A) Schematic overview of histone domains. (B) Stepwise assembly of the core histones into a functional nucleosome wrapped around DNA. Adapted from
1.2 The histone code and chromatin dynamics

Within the context of gene regulation, the most interesting structural feature of nucleosomes is probably the “histone tails”. These are intrinsically disordered and highly conserved domains located at the N-terminal of all core histones and also at the C-terminal of H2A and H2B. Histone tails can be decorated with a variety of post-translational modifications (PTMs) that serve as binding platforms for the recruitment of proteins with PTM-recognition domains (Strahl and Allis 2000; Jenuwein and Allis 2001) (Figure 1-3). These latter include several types such as the Bromodomains (binds acetylated histones), Chromodomains and PHD fingers (binds methylated histones), and they are commonly referred to as “histone readers” (Marmorstein and Berger 2001; Maurer-Stroh et al. 2003; Kim et al. 2006; Champagne and Kutateladze 2009; Yun et al. 2011). When readers bind to modified histones, they can exert a variety of activities that will ultimately lead to changes in gene expression.

The best characterized PTMs are methylation and acetylation of lysine (K) and arginine (R) residues imposed by dedicated enzymes like histone methyltransferases (HMT) and histone acetyltransferases (HAT), denominated as “writers”. Likewise, histone PTMs can be removed by “erasers” such as histone demethylases (HDMTs) and histone deacetylases (HDACs). The fact that alone or in combination, histone PTMs can affect gene expression, is known as “the histone code” (Strahl and Allis 2000; Jenuwein and Allis 2001). This represents a layer of gene regulation additional to the DNA sequence and allows to organize the chromatin into active or repressive regions. For example, the tri-methylation of histone 3 at lysine 4 (H3K4me3) or the acetylation of the same histone at lysine 27 (H3K27ac), mark the beginning of actively transcribed genes, while the tri-methylation of histone 3 at lysine 27 (H3K27ac) is preferentially found at repressed promoters (Schneider et al. 2004; Bernstein et al. 2005).
Figure 1-3. Overview of common histone PTMs
Tail sequences belong to human canonical histones. Taken from (Anon n.d.)

Figure 1-4. Consequences of ATP-dependent chromatin remodelling
(A-C) Cartoon depicting the main tasks carried out by ATP-dependent chromatin remodelers. Taken from (Becker and Workman 2013)
Chapter 1

Notably, while PTMs can by themselves alleviate or strengthen DNA-histone interactions, the ultimate purpose of the histone code is to recruit specialized proteins. These can be transcription factors required for gene expression or chromatin remodelers, which alter the chromatin state to actively promote or repress transcription. Chromatin remodelers are able to mobilize nucleosomes by breaking histone-DNA contacts through ATP hydrolysis (Längst and Manelyte 2015; Clapier et al. 2017). They carry three major tasks: 1) chromatin accessibility, through nucleosome eviction, sliding or histone exchange (Figure 1-4A); 2) chromatin assembly, aided by chaperones to ensure deposition of nucleosomes (Figure 1-4B); and 3) maintenance of a correct nucleosome spacing (Figure 1-4C). Chromatin remodeling represent the first step towards gene activation as regulatory DNA-sequences need to be exposed to be accessible by transcription factors.

1.3 RNA polymerase II transcription

Polymerases are essential enzymes for all living organisms. Highly conserved through evolution, they perform the synthesis of the RNA molecules within the cell nucleus. In bacteria, one polymerase is in charge of producing all RNAs, while in eukaryotes, three dedicated polymerases (Pol I, Pol II and Pol III) are responsible for the synthesis of the different classes of RNA. Eukaryotic RNA polymerases were initially discovered by Bob Roeder and William Butter in 1969 using chromatographic separation of rat liver nuclei and developing sea urchin embryos (Roeder and Rutter 1969). Later work by the laboratory of Bob Roeder identified the target genes of all three polymerases, being Pol I responsible for the synthesis of ribosomal RNAs while Pol II and Pol III were respectively involved in the synthesis of messenger RNAs (protein-coding) and the 5S and transference RNAs (Weinmann and Roeder 1974; Weinmann et al. 1974). These findings were pivotal for the field of transcription, standing as the foundation for many later structural, genetic and imaging studies.
Transcription of protein-coding genes by Pol II is with no doubt the most extensively studied. In comparison with other classes of RNA, the production of mRNA is tightly regulated. This is probably due to 1) the fundamental role of proteins in most cellular activities and 2) their requirement for differential tissue-specific expression. To better understand the process of mRNA synthesis, a lot of effort has been put into the functional and structural characterization of the Pol II.

Pol II is a large complex (~ 550 KDa) composed of 12 subunits highly conserved through evolution. The three-dimensional structure of Pol II adopts a “crab-claw” shape that efficiently engages single-stranded DNA and allows mRNA synthesis (Cramer et al. 2001; Klug 2001; Woychik and Hampsey 2002; Hahn 2004; Bernecky et al. 2016) (Figure 1-5A). These are generally named using the prefix “Rpb”, followed by increasing numbers that represent the protein size. A key feature for the regulation of the transcription cycle lies at the C-terminal domain (CTD) of Rpb1, the largest of Pol II subunits. This unusual domain is intrinsically disordered and consists of multiple repeats of the heptapeptide “YSPTSPS” sequence. While the number of these repeats can be variable across species, the CTD itself is highly conserved through evolution and is essential for life. In fact, deletion of most repeats or introduction of external sequences is lethal in yeast and mouse (Nonet et al. 1987; Bartolomei et al. 1988; Zehring et al. 1988; Allison et al. 1988).

The function of the CTD is to serve as a scaffold for additional transcription factors involved in transcriptional regulation (Thomas et al. 1998; Hsin and Manley 2012; Chen et al. 2018; Schier and Taatjes 2020). This mainly occurs through changes in the phosphorylation status of its serine 5 and 2 residues (Ser5 and Ser2 (Figure 1-5B). For example, when Pol II is initially recruited to the TSS of a gene, its CTD is completely unphosphorylated. This attracts a huge multisubunit complex called Mediator, important for the crosstalk between the core transcription machinery and gene-specific activators (Lu et al. 1991; Myers et al. 1998). By the time the transcription machinery is fully assembled, the
levels of Ser5 phosphorylation (Ser5-P) are high, which decreases the affinity for the Mediator complex, recruits capping enzymes and facilitates the release of Pol II into the productive elongation (Cho et al. 1997; McCracken et al. 1997; Jeronimo and Robert 2014; Wong et al. 2014). As Pol II makes its way out of the TSS, the levels of Ser5-P decrease while those of Ser2-P accumulate along the gene body (Komarnitsky et al. 2000), reaching its maximum at the end of gene. Unlike Ser5-P, Ser2-P increases Pol II affinity for elongation and splicing factors required for further the processing of the mRNA as transcription occurs.

When Pol II has finished transcribing a gene, it returns to its hypophosphorylated state by the action of specific phosphatases and is then released from the DNA to begin a new round of transcription. Importantly, although the CTD is essential in vivo, is not required for basal transcription in vitro, supporting its role as a platform that integrates signals from specific factors to finely coordinate transcription (Zehring et al. 1988; Kim and Dahmus 1989; Buratowski and Sharp 1990).
Figure 1-5. Pol II and the CTD phosphorylation cycle

(A) The structure of the RNA polymerase II (Pol II) resembles a “crab-claw” and is highly conserved through evolution. Due to the inherent disorder of the CTD it cannot be seen in this structural model. Data was obtained from protein data bank under accession number 5IY6. (B) Pol II arrives at the promoter without CTD phosphorylation. After Pol II recruitment, Pol II CTD can be phosphorylated at serine residues at position 2 and 5 (Ser2 and Ser5). Ser5 phosphorylation predominates at the beginning of the gene, whereas Ser2 phosphorylation is enriched at the end of a gene.
1.4 Basal transcription factors are required for accurate Pol II initiation

Despite its ability for DNA-dependent RNA synthesis, Pol II alone has a low affinity for double-stranded DNA. This became apparent from early transcription studies, in which accurate Pol II transcription was only possible in the presence of several accessory components known as basal or general transcription factors (BTFs/GTFs) (Weil et al. 1979; Matsui et al. 1980). Technically, the name BTFs is more appropriate as it reflects the capacity of these factors to sustain basal transcription in vitro systems without the presence of activators. Besides, while BTFs are required for expression of nearly all Pol II genes, accumulating evidence suggests that some of these factors can be replaced by non-canonical analogs at certain genes and in a tissue-specific manner during development (Hochheimer and Tjian 2003; Goodrich and Tjian 2010; Akhtar and Veenstra 2011; Kedmi et al. 2014; Duttke 2015), arguing against a general or universal transcription machinery.

The BTFs include six RNA Pol II dependent transcription factors (TFII) named TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH. These factors are able to recruit Pol II at the start site of genes, change its CTD phosphorylation status and induce melting of the double-stranded DNA to form the transcription bubble (Schier and Taatjes 2020). Together, Pol II and BTFs assemble into a massive transcriptional machinery called the pre-initiation complex (PIC) (Figure 1-6B). Except for TFIIB, all PIC components are made of several subunits reaching up to a total size of 2.4 megadaltons (MDa) when the full complex is formed (Murakami et al. 2015). This process takes place in several steps and occurs at the region surrounding the TSS of genes, referred to as the core promoter (Figure 1-6A). Importantly, most of our mechanistic understanding about transcription initiation comes from in vitro experiments, the majority of which are based on promoters containing the TATA-box, a well-known regulatory sequence located ~ 30 bp upstream of the TSS. TBP, a component of the TFIID complex, has a strong affinity for the TATA-box and their
interaction has been traditionally considered as the trigger for PIC formation (Roeder 1996; Woychik and Hampsey 2002; Hahn 2004).

The contacts between TBP and the TATA-box are stabilized by TFIIA, a dimer composed of GTF2A1 and GTF2A2 (Tan et al. 1996; Imbalzano, Kwon, et al. 1994; Imbalzano, Zaret, et al. 1994; Kang et al. 1995; Ozer et al. 1998). TFIIA binds to TBP and shields its interaction surface from negative regulators like Nc2 and Mot1, which otherwise would promote its dissociation from DNA (Kamada et al. 2001; Klejman et al. 2004; van Werven et al. 2008; de Graaf et al. 2010; Wollmann et al. 2011; Spedale et al. 2012; Koster and Timmers 2015). Besides, the TFIIA-TBP-DNA complex is further stabilized by GTF2A2 interactions with the promoter region right upstream of the TATA-box (Tan et al. 1996; Lagrange et al. 1996; Bleichenbacher et al. 2003; Shao and Zeitlinger 2017). Early studies about the requirement of TFIIA for PIC assembly were controversial, with results ranging from TFIIA having no impact on basal transcription to being indispensable (Sawadogo and Sentenac 1990). However, it was shown later that apart from assisting TBP-DNA interactions, TFIIA can facilitate or antagonize the recruitment of transcriptional regulators (Ge and Roeder 1994; Auble et al. 1994; Sun et al. 1994; Kobayashi et al. 1995; Papai et al. 2010), suggesting that initial differences in activity were probably due to the absence of TFIIA-interacting factors in highly-purified systems.
TFIIB is the third factor to join the PIC. It further stabilizes the TBP-TFIIA complex by directly interacting with TBP as well as with DNA right upstream and downstream of the TATA-box, which is possible due to a TBP-induced 90 degrees promoter bend around the TATA-box (J. L. Kim et al. 1993; Lee and Hahn 1995; Nikolov et al. 1995; Lagrange et al. 1998; Wolner and Gralla 2001; Zhang et al. 2002; Deng and Roberts 2005). Besides, TFIIB is important for TSS selection (Pinto et al. 1992; Berroteran et al. 1994; Lee and Hahn 1995; Zhang et al. 2002) and provides a binding platform for the concomitant recruitment of the RNA Pol II (Murakami et al. 2015; He et al. 2016).

Pol II lands at the forming PIC by the hand of TFIIF, a dimer composed of GTF2F1 and GTF2F2 (Sopta et al. 1985; Conaway et al. 1987; Price et al. 1987; Flores et al. 1990; Flores et al. 1991). TFIIF plays several important roles for transcription initiation. (1) The strong affinity between TFIIF and Pol II facilitates its interaction with the TFIIA-TBP-TFIIB-DNA complex (Sopta et al. 1985; Price et al. 1987; Flores et al. 1991; Chung et al. 2003; Chen et al. 2010); (2) its recruitment exposes additional promoter-DNA contact surfaces that provide stability to the early PIC (Conaway et al. 1991; McCracken and Greenblatt 1991; Buratowski et al. 1991; Aso et al. 1994; Robert et al. 1998); (3) shows a predominant role in TSS selection (Sun and Hampsey 1995; Yan et al. 1999; Ghazy et al. 2004; Khaperskyy et al. 2008) and (4) promotes TFIIE and TFIIH recruitment via direct interactions (Flores et al. 1990; Murakami et al. 2015; Schilbach et al. 2017; Greber and Nogales 2019).

TFIIE and TFIIH work cooperatively and both are essential for the formation of a transcriptionally competent Pol II. TFIIE participates in the formation of a structural bridge that keeps DNA close to the Pol II cleft and stimulates TFIIH enzymatic activities (Goodrich and Tjian 1994; Ohkuma and Roeder 1994; Ohkuma et al. 1995; Serizawa et al. 1995; Holstege et al. 1996; Okamoto et al. 1998; Greber and Nogales 2019; Schier and Taatjes 2020). On his side, the XPB translocase subunit of TFIIH pushes downstream DNA into the
Pol II catalytic site, and because the upstream promoter is firmly engaged through TBP-TFIIB-TFIIE interactions, such backtracking induces a torsional strain that leads to the unwinding of the double stranded-DNA and the formation of the so-called transcription bubble (Kim et al. 2000; Grünberg et al. 2012; He et al. 2013; He et al. 2016). Besides, XPB activity requires major structural rearrangements that involve the repositioning of the kinase CDK7 subunit of TFIIH near the Pol II CTD. This results in the phosphorylation of Ser5 residues, which promotes elongation by releasing Pol II from the PIC contacts (He et al. 2013; Schier and Taatjes 2020).

**Figure 1-6. Stepwise assembly and structure of the PIC**

(A) First, TFIID is recruited to core promoter via interaction between TBP and the TATA-box, which results in a 90 degrees bending of DNA. TBP-TATA interactions are stabilized by TFIIB and later TFIIA. Pol II is then recruited and engages the TSS stabilizing the early PIC together with TFIIF. Lastly, the PIC is complete by the incorporation of TFIIE and TFIIH, which “open-up” the double-stranded DNA and create the transcription bubble. (B) Structural model of a partial human PIC at the super core promoter (SCP, see below). Adapted from (Shao and Zeitlinger 2017).
1.5 Pol II pausing, a rate-limiting step for metazoan gene transcription

Pol II recruitment has been traditionally believed to be the most rate-limiting step of transcription. This idea comes from early in vitro systems mainly based on yeast transcription, in which the formation of the PIC was sufficient to drive Pol II mRNA synthesis. However, several studies published during the 1970s and early 80s revealed that such systems often did not result in the expression of transcripts, suggesting the need for other factors that regulate Pol II after initiation (Fraser et al. 1978; Gariglio et al. 1981). Later experiments using the inducible promoter Hsp70 showed that transcriptionally engaged Pol II accumulates right downstream of the TSS, associated with short-nascent RNA transcripts that vary between 20-60 nt in size (Rougvie and Lis 1988; Rasmussen and Lis 1993). This phenomenon was termed as “pausing”, for its resemblance with the paused polymerase found in bacteria by Jeffrey Roberts in 1985. Following studies during the 90s in fly and mammals showed that pausing occurred at different types of genes (Rougvie and Lis 1990; Strobl and Eick 1992; Krumm et al. 1992; Plet et al. 1995; Law et al. 1998), and more recently, the development of techniques that can detect DNA-bound Pol II genome-wide has revealed that Pol II pausing is a widespread phenomenon in metazoans (Muse et al. 2007; Core et al. 2008; Nechaev et al. 2010).

The establishment and release of Pol II pausing is essentially regulated by three different factors: DSIF, NELF and P-TEFb (Figure 1-7A). First, DSIF and NELF bind and stall the RNA polymerase II after it has transcribed at least 20-30 nt (Missra and Gilmour 2010). Such stalling is mainly induced by the action of NELF, which engages the Pol II funnel, restricting the polymerase mobility and creating a tilted RNA-DNA hybrid that is incompatible with the addition of new RNA nucleotides (Vos et al. 2018). This repressive state is overcome by the kinase activity of P-TEFb, which upon recruitment phosphorylates DSIF and NELF, causing NELF to dissociate from the promoter and turning DSIF into a
positive factor that promotes elongation. P-TEFb also phosphorylates the Pol II CTD at Ser2, which serves as a binding platform for transcription elongation factors and marks the beginning of productive elongation.

Pausing has been proposed to serve as a mechanism that primes genes for rapid activation (Zeitlinger, Stark, et al. 2007; Guenther et al. 2007). This seems to particularly important during development, where the formation of tissues requires coordinated expression of specific genes across cell populations (Boettiger and Levine 2009; Lagha et al. 2013). It appears that by remaining bound to the promoter region, the paused Pol II prevents binding of nucleosomes and helps maintain an active chromatin state. Presumably, this may help to bypass potentially slow steps like promoter opening occurring during early gene activation. In agreement with this, several studies have shown that reducing the levels of pausing factors leads to an increased nucleosome occupancy in the core promoter region and a reduction in the expression levels of many genes (Gilchrist et al. 2008; Gilchrist et al. 2010).
Notably, all genes experience pausing to some degree as blocking Pol II pause-release using transcription inhibitors like flavopiridol (FP) or 5,6-dichloro-1-β-D-ribofuranosylbenzimidazole (DRB) has been shown to affect the expression of nearly all Pol II genes in *Drosophila* and mammalian cells (Chao and Price 2001; Ni et al. 2008; Rahl et al. 2010; Henriques et al. 2013; Jonkers et al. 2014), indicating that pausing is an obligatory step of transcription rather than a special feature of developmental promoters. However, the stability of the paused Pol II is widely variable and can range from tens of seconds to tens of minutes. The nature of this variability is highly gene specific and while increased activation can modulate the stability of pausing by promoting release of Pol II into elongation (Min et al. 2011; Danko et al. 2013; Buckley et al. 2014), to a large extent, the degree of pausing seems to be determined by the core promoter sequence. In *Drosophila*, for instance, promoters that experience stable pausing are enriched in specific DNA-sequences such Inr, DPE or PB, whereas those with short Pol II pausing either contain TATA-box or lack pausing motifs (Hendrix et al. 2008; Gaertner et al. 2012; K. Chen et al. 2013; Zeitlinger, Zinzen, et al. 2007; Shao et al. 2019).

**Figure 1-7. Schematic overview of the establishment of Pol II pausing and release**

(A) After the pre-initiation complex has formed, Pol II transcribes about 20-60 nt before it is stalled by the action of DSIF and NELF. After, P-TEFb is recruited and phosphorylates. NELF, DSIF and the Ser2 of Pol II CTD, releasing NELF from the pause complex and transforming DSIF into a positive factor. In addition, Ser2 phosphorylation marks the beginning of elongation and will serve as a binding platform for additional transcription factors.
1.6 The core promoter region: at the heart of transcription

A core promoter can be defined the minimal stretch of DNA required to assemble the PIC and is often referred to as the “gateway” to transcription, since all signals leading to Pol II initiation must eventually converge at this region (Juven-Gershon, Hsu, Theisen, et al. 2008). Early studies in bacteria first identified the core promoter as small DNA segments indispensable for the expression of structural genes (Ippeen et al. 1968). Later studies in eukaryotic transcription started to reveal that specific sequences or “elements” within the core promoter region were responsible for the accurate positioning of the BTFs and the Pol II near the initiation site (Weil et al. 1979; Manley et al. 1980). Hence, the core promoter is itself a regulatory sequence.

Initially, core promoters were thought to be static and generic entities, with little role on transcription apart from directing Pol II binding. This notion roots in the early transcriptional studies, where most assays employed a small number of strong viral or metazoan promoters sharing a common feature – the presence of a TATA-box (Vo Ngoc, Wang, et al. 2017). This paradigm has certainly changed with the availability of genome-wide analysis. These have exposed that core promoters are extremely diverse in sequence composition and that no universal core promoter element exists in vivo. Besides, within the “palette” of core promoter possibilities, those containing a consensus TATA-box represent a minority, with less than 10% of fly and human promoters containing this element, and only ~20% in yeast (Basehoar et al. 2004; Lenhard et al. 2012; Koster et al. 2015). Notably, despite the striking diversity in sequence composition, many specific DNA-sequences have recurrently been found within core promoters over the past years using experimental techniques and computational approaches. A table summarizing their main features is
presented in table 1-1 and figure 1-8, and a brief description of the most studied elements is given below.

The TATA-box was the first core promoter element found in eukaryotes and it is overall the most extensively studied. Its discovery was the result of a comparative analysis of the upstream sequences of the H2A gene in various species, revealing a conserved enrichment of A-T nucleotides ~ 25-30 bp upstream of TSSs (Lifton et al. 1978; Breathnach and Chambon 1981), whose importance for transcription was later demonstrated by mutational analysis (Grosschedl et al. 1981; Mathis and Chambon 1981; Giangrande et al. 1989; Ponjavic et al. 2006). As discussed above, the TATA-box is recognized by TBP and both early evolutionary inventions as they are not only present in eukaryotes, but also in archaea (Blombach et al. 2016). The consensus sequence for a strong TATA-box is considered STATAWAWR, however, many TATA-like sequences carrying one or two mismatches from the consensus have been described to be important for transcription initiation and therefore a less strict TATAWR consensus sequence has been recently proposed (Vo Ngoc, Wang, et al. 2017).

TATA elements are sometimes flanked by small sequences termed the BREu and BRED (Vo Ngoc et al. 2019). The BREu is located right upstream of the TATA-box while the BRED is right downstream and their derived consensus sequences are SRCGCC and RTDKKKK, respectively (Lagrange et al. 1998; Deng and Roberts 2005). The function of BREs elements remains controversial, as they have been shown to affect transcription positively and negatively (Lagrange et al. 1998; Evans et al. 2001; Deng and Roberts 2005; Juven-Gershon, Hsu and Kadonaga 2008). However, it is important to note that these elements have not been extensively studied and such consensus sequences have not been validated with more recent methodology. Besides, BREs are not recognized by TFIID, but TFIIB, and because promoter binding of TFIIB depends on prior TATA-TBP interactions
(see section 1.3), the presence of TATA-box is a prerequisite for the proper functioning of these sequences.

Another well-studied core promoter element is the initiator (Inr), which directly encompasses the TSS of high proportion of genes. The Inr can work synergistically with the TATA-box (Emami et al. 1997), but it is much more abundant than this latter (Ohler et al. 2002; Vo Ngoc, Cassidy, et al. 2017). The consensus sequence for the Inr has higher information content in flies than in human, being TCA_{+1}KTY and CA_{+1}, respectively, and where transcription initiates at the A_{+1} nucleotide (Vo Ngoc, Wang, et al. 2017). In the absence of TATA, the Inr element is most often associated with a strong GC-rich sequence downstream of the TSS called the Downstream Promoter Element (DPE). The DPE was initially discovered in Drosophila (Burke and Kadonaga 1996) and was thought to be rare in mammals (Burke and Kadonaga 1997; Kutach and Kadonaga 2000). However, a very recent study by Ngoc and Kadonaga using high-throughput analysis of randomized promoter elements showed that the DPE was widely used also in human promoters (Vo Ngoc et al. 2020). Notably, several TAFs has been shown to specifically target the Inr and MTE/DPE sequences and strict spacing between these elements appears to be required for efficient binding of the TFIID complex (Burke and Kadonaga 1997; Louder et al. 2016).

Apart from the DPE, other two GC-rich sequences have been found to be overrepresented in the downstream core promoter region: The Motif Ten element (MTE) and the Pause Button (PB). The MTE was discovered through computational analysis (Ohler et al. 2002) and later found to work cooperatively with the Inr and DPE to enhance TFIID binding (Lim et al. 2004; Theisen et al. 2010). On the other hand, the PB was derived from analysis at promoters where Pol II pausing was stable and was often located in the position of the DPE (Hendrix et al. 2008). Notably, the PB and DPE are similar in sequence
composition and in both cases, the association with the Inr seems to be an intrinsic feature of these elements.

An aspect worth mentioning is that there appears to be a clear duality between TATA- and MTE/DPE/PB-containing promoters, as they rarely occur together in the same core promoter in vivo (Burke and Kadonaga 1996; Butler and Kadonaga 2001; Shao et al. 2019; Vo Ngoc et al. 2020). In this context, it is interesting to note that TBP alone can directly recognize the TATA element in vitro and support basal transcription, while TATA-less promoters containing Inr, MTE or DPE show a strong dependency on TAFs for activity (Wright et al. 2006; Donczew and Hahn 2018), suggesting different mechanisms for TFIID recruitment and transcription initiation depending on core promoter sequence composition. Moreover, TBP was proposed to activate TATA-promoters but repress those containing DPE through a specific regulatory circuit controlled by Nc2 and Mot1, which regulate the TBP DNA-binding dynamics (Hsu et al. 2008). However, it was later demonstrated that such effect was likely due to competition for TBP between the intrinsically-preferred TATA-containing promoters and the more disfavored TATA-less promoters (Zentner and Henikoff 2013). Besides, the existence of a functional synthetic super core promoter (SCP) containing TATA, Inr and MTE/DPE (Juven-Gershon et al. 2006), and which efficiently binds TFIID in vitro (Cianfrocco et al. 2013; Louder et al. 2016; Patel et al. 2018), suggests that these elements do not necessarily antagonize each other and can function together.

In flies, additional motifs have been found using computational analysis of large collections core promoters (Ohler et al. 2002; FitzGerald et al. 2006). These include the Ohler motifs 1, 6 and 7 and DNA replication-related element (DRE). An important feature of these elements is that do not seem to require a strict positioning relative to the TSS, as occurs with the TATA, Inr and MTE/DPE/PB. Furthermore, while these elements are located within the core promoter region, they appear to be recognized by specific TFs other than TFIID. For example, the Motif1 Binding Protein (M1BP) directly binds Ohler 1 motif (Li
and Gilmour 2013; Baumann and Gilmour 2017) and the DNA replication-related element factor (DREF) recognizes the DRE (Hirose et al. 1993). It is possible that these elements promote transcription through activator-dependent interactions with TFIID rather than directly stabilizing its DNA binding.

Another well-studied core promoter motif is the polypyrimidine initiator, also known as the TCT element. The TCT motif essentially works as an initiator as it also encompasses the TSS (Vo Ngoc, Wang, et al. 2017). Its consensus sequence is YYC_{+1}TTTYY in flies (Parry et al. 2010) and YC_{+1}TYTYY in humans (Parry et al. 2010; Vo Ngoc, Cassidy, et al. 2017), where transcription initiates at the C_{+1} nucleotide instead of the A_{+1}, as occurs with the canonical Inr. The TCT element is extremely rare, yet highly relevant, as it is dedicated almost exclusively to the synthesis of Ribosomal Protein Genes (RPG) in both flies and humans (Parry et al. 2010). Another peculiar aspect about the TCT promoters, at least in flies, is that transcription appears to be dependent on the TBP-related factor 2 (TRF2) instead of the canonical TBP (Y.-L. Wang et al. 2014). However, other TAFs have been shown to be important for RPG transcription, leading to propose that TRF2 could substitute TBP within TFIID depending on the core promoter composition (Baumann and Gilmour 2017).

Finally, it is important to highlight that the presence of different core promoter elements is intrinsically related to their function and patterns of transcription initiation (Figure 1-8) (Haberle and Stark 2018). For example, at promoters that contain TATA, Inr, MTE, or DPE, Pol II initiates transcription from one or few nucleotides in close proximity to each other (<= 6 nt) (Roider et al. 2009; Rach et al. 2011). This is referred to as “focused”, “narrow” or “sharp” transcription and is associated with the expression of developmental and tissue-specific genes, which are regulated in a spatiotemporal manner (Rach et al. 2011). By contrast, dispersed promoters are linked to ubiquitously expressed housekeeping genes and show a “dispersed” or “broad” pattern of initiation, with several TSSs distributed across a window of 50-100 nt (Roider et al. 2009; Rach et al. 2011). In vertebrates, dispersed
promoters are associated with the presence of CpG islands (Carninci et al. 2006) while in Drosophila these have well-defined motifs such as DRE, Ohler1, Oherl6 and Ohler7, discussed above (Rach et al. 2009).

Core promoters differ in sequence composition, initiation pattern and function

Core promoters encompass the region immediately upstream and downstream of the TSS and represent the minimum space necessary to assemble the PIC on DNA. Core promoters possess different features, but can often be grouped based on sequence composition, initiation pattern, type of target gene they are found at or even nucleosome occupancy. For example, focused promoters utilize one a few strong TSSs that occur near each other, are enriched for elements like TATA, Inr, MTE, DPE and PB, are typically found at stimulus-response tissue-specific genes and show little or none positioning of the +1 nucleosome. In contrast, at dispersed promoters initiation starts at several places within a broad window of 100-150 bp, they are enriched for elements like DRE, Motif1, Motif6 and Motif7, encompass the expression ubiquitously expressed housekeeping genes and show strong positioning of the +1 nucleosome. An exception to this classification are promoters enriched in the TCT element, these present a focused pattern of initiation despite being found at nearly all ribosomal protein genes (RPG), which are strongly transcribed at all tissues and times of development.
<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
<th>Position</th>
<th>Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>TATA box</td>
<td>TATAWAWR</td>
<td>−35 to −25</td>
<td>The first eukaryotic core promoter element. Binds TBP.</td>
</tr>
</tbody>
</table>
| TFIIB recognition element (BRE) | BRE<sup>-</sup>: SSRCGCC BRE<sup>+</sup>: RTDKKK | −42 to −35  
  −23 to −17 | GC-rich sequences flanking the TATA-box that stabilize TFIIB binding.                                                                     |
| Initiator (Inr)               | TCAKTY                   | −2 to +4     | The most frequent core promoter motif. Encompasses the TSS and binds TFII D.                                                               |
| polypyrimidine Inr (TCT)      | YYCTTTYY                | −2 to +6     | Inr variant specifically dedicated to the synthesis of ribosomal protein genes. Binds the TBP-related factor 2 (TRF2).                    |
| Downstream promoter element (DPE) | KCGGTTSK                | +25 to +35   | GC-rich downstream sequence important for TFII D binding at TATA-less promoters. Several variants found. Works cooperatively with the Inr. |
| Motif ten element (MTE)       | CSARCSSA                 | +18 to +24   | Downstream sequence that functions cooperatively with DPE and Inr in TFII D binding. Not as well characterized as other motifs.          |
| Pause button (PB)             | KCGRWCG                  | −5 to +55    | GC-rich downstream sequence similar to the DPE. Enriched at promoters with stably paused Pol II.                                           |
| DNA replication-related element (DRE) | WATCGATW              | −100 to 0    | First identified as important for high expression of cell-proliferation genes. Binds DREF and can work cooperatively with the TCT. Occurs upstream without strict positioning relative to the TSS. |
| Ohler 1                       | YGGTCACACTR              | −100 to 50   | Enriched at dispersed housekeeping promoters. Can be found upstream of the promoter but also encompassing TSSs. Binds M1BP and can work cooperatively with Ohler 6 and TCT. |
| Ohler 6                       | YRGATWTWTTY              | −150 to 25   | A-T rich sequence over-represented at dispersed housekeeping promoters. Often found upstream of the Ohler 1.                                |
| Ohler 7                       | CAKCNCTR                 | −100 to 50   | Enriched upstream of dispersed housekeeping promoters.                                                                                   |
1.7 The TFIID complex is a specialized core promoter recognition machine

Together with the other basal transcription factors, TFIID was first identified from HeLa cell extracts using cation-exchange phosphocellulose chromatography in 1974. This method allowed separation of DNA-binding proteins using increasing amounts of salt concentration and resulted in four-differently eluted nuclear fractions named from A to D. While most PIC components were found in fractions A-C, TFIID was located in fraction D (transcription factor of Pol II, fraction D), strongly suggesting a role in DNA-binding and therefore in core promoter recognition. Later DNase I footprinting experiments demonstrated the ability of TFIID to interact with the TATA-box and emphasized its pioneer role in nucleating the PIC.

During that time, the molecular size estimates of TFIID ranged between 140-1300 kDa (Timmers and Sharp 1991). However, experiments using yeast extracts led to the purification of a smaller protein (~27 kDa) in the D fraction, which presented TATA-binding properties and had the ability to drive transcription from in vitro reconstituted systems, even when using TFIID-depleted human extracts. The presence of conserved repeated domains within yeast TFIID facilitated the identification of the cDNA encoding for human TFIID. The surprise came as the predicted size of the cloned human TFIID was only about 38 kDa. Moreover, while partially purified TFIID from D fractions led to protection of core promoter sequences from -45 to +35, the 38-kDa TFIID only protected the region around the TATA-box.

These data suggested that the recombinant 38-kDa human TFIID, today widely known as the TATA-binding protein (TBP), was part of a bigger TFIID complex containing other subunits. This was in line with the fact that TBP alone was able to support basal levels of transcription in vitro, but not capable of responding to activator signals, suggesting that other TBP-associated factors (TAFs) were indeed required for communicating the PIC with
distal regulatory elements. Over the years, extensive biochemical and structural studies have revealed the existence of 13 different TAFs, of which 6 (TAF4, -5, -6, -9, -10 and -12) are present in two copies. Together with TBP and TAFs form the canonical holo-TFIID complex.

1.8 Composition and assembly of the TFIID complex

Most TFIID subunits are extremely conserved phylogenetically and can include splice variants, paralogs and shared components. Originally, TAFs were named based on their estimated molecular size in the species they were identified. However, this was confusing since the size of conserved TAFs orthologs was variable across species and therefore a unified nomenclature was proposed in 2002 (Tora 2002). In this section, I will briefly review the main structural and functional features of TBP and each TAF gathered over the years.

TBP

TBP was the first subunit of TFIID to be discovered and it has been extensively characterized (Horikoshi et al. 1989; Hahn et al. 1989; Kao et al. 1990). Structurally, TBP can be divided into two domains: 1) a highly conserved C-terminal core and 2) a less-conserved and flexible N-terminal domain. While this latter is variable in size and sequence, it is conserved in vertebrates (Bondareva and Schmidt 2003). The functions of the N-terminal domain are not completely understood but is thought to module the DNA-binding capacity of TBP. In humans, this region contains a poly-gluatmine (polyQ) stretch with variable repeats and mutations affecting this number have been associated with distinct neurological disorders (Reid et al. 2004; Schaffar et al. 2004; Friedman et al. 2007; Huang et al. 2015; Yang et al. 2016). On the other hand, the core region of TBP is well characterized and extremely conserved among archea and eukaryotes (80% identity between yeast and mammals), suggesting an ancient role in transcription initiation (Akhtar and Veenstra 2011).
Its tridimensional structure adopts a quasi-symmetrical horse-saddle shape with a convex (top) and a concave (bottom) surface composed of a total of 4 alpha-helices and 10 β-sheets (Y. Kim et al. 1993; Nikolov et al. 1995; Geiger et al. 1996).

The underside concave part of TBP is made of eight curved β-sheets (four in each half of the saddle) rich in positively charged lysine and arginine residues that interact with the negatively charged phosphate backbone of DNA. Besides, the upper convex surface of TBP is built by four alpha helices (two in each side of the saddle and perpendicular to each other), a basic peptide linking both halves of the saddle and the remaining two β-sheets. This region of TBP is pivotal for gene expression as it serves as a binding platform for many proteins involved in transcription initiation including TAF1, TAF11/13, TFIIA, TFIIB, NC2 and BTAF1 (Kamada et al. 2001; Juo et al. 2003; Klejman et al. 2005; Wollmann et al. 2011; Anandapadamanaban et al. 2013; Knutson et al. 2014; Gupta et al. 2017). Notable, while in this thesis work I place most focus on transcription by the RNA Pol II, TBP is also required for the transcription of RNA Pol III and I genes, where the same surface interacts with BRF1 and SL1, respectively. In addition, it is worth mentioning that several proteins that interact with the convex surface of TBP such as TAF1, TAF11/13, NC2 and BTAF1 can also bind its concave part to regulate its DNA binding.

Unlike most transcription factors, TBP binds to the minor groove of DNA and causes a distortion in the DNA molecule that leads to a 90 degrees bend and partial unwind of double-helix (Y. Kim et al. 1993; Nikolov et al. 1995; Nikolov et al. 1996). At TATA-promoters this is achieved by two kinks exerted by two pairs of highly conserved and well positioned phenylalanines within the concave surface of TBP, which insert between the first T-A and between the 7th and 8th bp of the sequence. This causes a partial unwind of the 6 middle bp of the TATA-box that widens the minor groove in a way that fits the concave part of TBP in an energetically favorable manner. The minor groove bending caused by TBP facilitates promoter interactions by other basal transcription factors that contact DNA right
upstream and downstream of the TATA-box and which would not be possible on linear DNA.

For a long time it was unclear whether DNA bending by TBP was also a feature of TATA-less promoters. Mutational analysis have shown that single substitutions within the TATA-box sequence decreases TBP-DNA binding affinity, suggesting that TBP may exhibit lower affinity and different degree of bending for degenerate TATA-like sequences and TATA-less promoters (Millán-Pacheco et al. 2009; Tora and Timmers 2010). While the TAFs and other BTFs factors like TFIIA may compensate for the lower TBP binding affinity to TATA-less promoters, differential DNA bending depending on the underlying sequences TBP is bound has been proposed to function as a physical cataliser for the efficiency of TBP removal from DNA by its negative regulators NC2 and BTAF1 after PIC assembly (Tora and Timmers 2010). In line with this model, TATA-promoters have been shown to exhibit a higher rate of TBP turnover than TATA-less promoters (van Werven et al. 2009). Moreover, evolutionary and structural analysis have revealed a co-conservation between the presence of phenylalanines found in the concave surface of TBP with that of its negative regulators NC2 and BTAF1 (Butryn et al. 2015). However, recent high-resolution structural data of TFIID at endogenous promoters revealed similar promoter bending by TBP among TATA and TATA-less promoters, which argues against differential bending as a the key facilitator for TBP eviction by NC2/BTAF1 (Chen et al. 2021). Thus, the molecular basis for this phenomenon are not clear.

An interesting aspect about the evolution of transcription initiation is that while TBP is sufficient for PIC assembly of all gene classes in yeast, several TBP-related factors (TRF1, TRF2 and TRF3) can replace the canonical TBP in different contexts in higher eukaryotes (Reina and Hernandez 2007; Jacobi et al. 2007; Akhtar and Veenstra 2011). TRF1 and TRF3 are specific to insects and vertebrates, respectively, while TRF2 is present in all metazoans. TRF1 was the first of these paralogs to be identified (Crowley et al. 1993). It has a 64%
identity with the canonical TBP, can bind to the TATA-box and stimulate basal transcription in vitro (Holmes and Tjian 2000). It is highly expressed in male germ cells and during the development of the central nervous system. In addition, genome-wide ChIP-seq and depletion analyses in Drosophila have shown that TRF1 is mainly associated with RNA Pol III transcription in vivo, whereas in yeast and human, which do not possess this paralog, transcription of these genes is taken care by TBP (Schramm and Hernandez 2002; Isogai et al. 2007).

Unlike TRF1 and TRF3, TRF2 is present in all metazoans and is the more distant of all TBP paralogs (40% identity with the canonical TBP core domain) (Dantonel et al. 1999). It is also less conserved between species than TBP, with only ~ 45% identity between yeast and human. It lacks affinity for the TATA-box, but can interact with TFIIA and TFIIB, suggesting an alternative mechanism for transcription initiation independent of TBP and TATA (Moore et al. 1999; Teichmann et al. 1999; Rabenstein et al. 1999). TRF2 plays important roles in gene expression during early embryonic development and germ cell differentiation in different species. Besides, in fly, expression of ribosomal protein genes (RPG) containing the TCT element requires TRF2 instead of TBP (Y.-L. Wang et al. 2014). In addition, TRF2 is present at DPE-containing TATA-less promoters and physically interacts with DREF, the factor that binds to the DRE element (Hochheimer et al. 2002; Vo Ngoc et al. 2019).

Lastly, TRF3 is the most closely related of TBP paralogs with 95% identity with the core TBP domain. It binds to the TATA-box, interacts with TFIIA and TFIIB and supports basal transcription in vitro (Bártfai et al. 2004; Jallow et al. 2004). TRF3 is mainly enriched in oocytes where it replaces canonical TBP and TFIID as oocytes mature from primary to secondary follicles (Bártfai et al. 2004; Jallow et al. 2004; Gazdag et al. 2007; Akhtar and Veenstra 2009; Xiao et al. 2006; Yang et al. 2006; Gazdag et al. 2009). Consistently, depletion of TRF3 is associated with defect in folliculogenesis and female sterility in mice.
Moreover, a rapid switch from TRF3- to TBP-dominated transcription has been observed during early embryonic development in fish, frog and mice, although residual levels TRF3 are maintained at the early embryo.

**TAF1**

TAF1 is present in one copy and is one the best characterized subunits of the TFIID complex. It is also the largest, with an approximate molecular size of 250 kDa in humans, 230 kDa in fly and 130 kDa in yeast. Biochemical and structural studies have revealed a variety of domains within TAF1. These include a TAF1 N-terminal domain (TAND), involved in interactions with TBP; a winged helix (WH) and a zinc knuckle (ZnK) domain positioned in the middle of TAF1 and in charge of binding to the core promoter; and two tandem bromodomain (BrDs) located C-terminally that recognize acetylated histones (Jacobson et al. 2000; Anandapadamanaban et al. 2013; H. Wang et al. 2014; Curran et al. 2018). In addition, domains within TAF1 have been shown to posses HAT and kinase activity *in vitro* (Mizzen et al. 1996; Dikstein et al. 1996), but the relevance of these functions remain to be demonstrated *in vivo*.

The TAND domain of TAF1 is the connection point between TBP and the rest of the TFIID complex and plays a critical role keeping TBP away from spurious interactions with DNA and other factors when is not promoter-bound (Anandapadamanaban et al. 2013; Patel et al. 2018). It can be divided in two smaller domains: 1) the TAND1, which similar to BTAF1 occludes the concave DNA-binding surface of TBP through hydrophobic interactions that mimicry binding to the TATA-box, and 2) the TAND2, which posseses several negatively charged residues that contact a basic region within the convex surface of TBP and thus prevents it from TFIIA binding.

The characterization of the WH domain was fruit of a co-crystalization study of TAF1 and TAF7 and set the structural basis for the DNA-binding capabilities of TAF1 (H. Wang et al. 2014). Notably, while the WH was shown to bind various types of promoters...
indiscriminately, stronger association with DNA was observed when the Inr and DPE elements were present. These results were consistent with earlier biochemical studies suggesting selective anchoring of TFIID through TAF1 (and TAF2) at promoters containing these motifs (Chalkley and Verrijzer 1999) and have been further confirmed more recently by cryo-EM studies (Louder et al. 2016; Patel et al. 2018). The TAF1-TAF7 co-crystalization also revealed two TAF7-interaction domains flanking the WH, an alpha helical and a triple barrel, which formed a tridimensional pyramid-like structure upon TAF1-TAF7 association.

TAF1 also presents a ZnK domain involved in DNA binding (Curran et al. 2018). Similar to the WH, this domain shows preferential association at optimized promoters containing both Inr and DPE, but is able to support basal levels of transcription at promoters lacking specific sequences. This lack in DNA-binding specificity observed for the WH and ZnK domains is probably an important feature for transcription due to the sequence diversity of Pol II core promoters. Besides, the BrD domains within TAF1 are also in charge of promoter binding through recognition of acetylated histones (Jacobson et al. 2000), a common feature of transcriptionally active genes.

**TAF2**

TAF2 is the second largest component of TFIID, is present in one copy and is mainly known for its role in promoter recognition. Its structure consists of a large N-terminal domain with sequence homology to the M1 family of metallopeptidases and a short and unstructured C-terminal domain that is rich in lysines and histidines (Malkowska et al. 2013). The metallopeptidase domain (MPD) is enzymatically inactive within TAF2, but provides an interaction surface for DNA and other members of TFIID like TAF8 and TAF1 (Ohtsuki et al. 2010; Trowitzsch et al. 2015; Louder et al. 2016). On the other hand, The TAF2 C-
terminal domain has been predicted to bind DNA due its positive charges, but this function has not been demonstrated.

Early studies reported TAF2 and TAF1 as key factors required for efficient transcription of genes containing Inr and downstream sequences (Burke and Kadonaga 1997). These observations are supported by more recent cryo-EM structures where TAF2 was mapped at the downstream core promoter region, adjacent to TAF1 and making contacts with DNA at three distinct positions (Louder et al. 2016). Interestingly, promoter contacts were mapped to a specific region of the MPD domain and not to the predicted DNA-binding C-terminal region, indicating that additional promoter interactions via TAF2 may occur in vivo.

TAF2 can also form a stable subcomplex with TAF8-TAF10 in the cytoplasm, where TAF2-TAF8 interactions are required for the nuclear import of TAF2 and its incorporation within the holo-TFIID complex (Trowitzsch et al. 2015). Notably, TFIID complexes lacking TAF2 has also been purified (Papai et al. 2009). While this may be due to substoichiometric presence or redundancy of TAF2 within TFIID, the discovery of partial complexes such as the trimeric TAF2-TAF8-TAF10 have led to propose that TFIID subcomplexes may exist as independent entities in the nucleus, presumably to dictate transcription from specific genes. However, promoter-specific recruitment or regulation by TFIID subcomplexes has not yet been demonstrated.

**TAF3**

TAF3 is present in one copy and is one the histone-fold domain (HFD) containing TAFs within TFIID. Its molecular size varies significantly between yeast, ~ 47 kDa, and higher eukaryotes, ~ 140 kDa. This is due to plant homeodomain (PHD) finger located in the C-terminal region of TAF3 that is not present int yeast (Malkowska et al. 2013). On the other hand, the HFD is located in the N-terminal part of the protein and it is phylogenetically conserved. TAF3-HFD specifically dimerizes with TAF10, its interaction partner within
TFIID, although extensive contacts with TAF1, TAF2 and TAF7 have also been reported by crosslinking mass spectrometry (CXMS) analysis in yTFIID (Birck et al. 1998; Kolesnikova et al. 2018).

The PHD domain of TAF3 is important for promoter recognition, as it was shown to selectively bind H3K4me3 in vivo (Vermeulen et al. 2007). Consistently, reduction in the levels of this histone mark resulted in a decreased in TFIID binding and expression from a subset of genes, providing a functional link between this modification and activated transcription (Lauberth et al. 2013). Besides, TFIID binding to H3K4me3 is inhibited by asymmetric dimethylation of H3R2, whereas H3K9 and H3K14 acetylation enhances its binding, which reveals an important crosstalk between TFIID and histone modifications during transcription initiation.

Outside of TFIID, TAF3 has been implicated in the regulation of gene-specific programs during early embryonic development and during differentiation form myoblasts to miotubules, where it associates with the TBP-paralog TRF3 (Deato and Tjian 2007; Deato et al. 2008; Hart et al. 2009). In both cases, expression of TAF3 and TRF3 was shown to be upregulated, while the levels of TBP-TFIID were greatly reduced, suggesting that canonical TFIID was not required for gene expression in such contexts. Surprisingly, later genetic studies showed that TRF3 null mice did not compromise adult myogenesis or tissue regeneration and further experiments demonstrated that TRF3 expression was actually absent in primary and established muscle cells (Malecova et al. 2016). Importantly, while these studies also observed TFIID subunits and TBP being downregulated during myoblast differentiation, reduced amounts of these proteins formed a detectable complex on promoters of muscle genes that was essential for their expression. Therefore arguing against a replacement of the canonical TFIID by a tissue-specific TRF3-TAF3 subcomplex.

TAF4
TAF4 is present in two copies and it contains a HFD on its C-terminal region, which uses to specifically partner with the HFD of TAF12 through an interaction that structurally resembles that of H2A and H2B (Bieniossek et al. 2013). Together, TAF4 and TAF12 has been proposed to be part of a nuclear TFIID core (cTFIID) complex that is essential for the assembly and integrity of the holo-complex.

Another key feature within TAF4 structure is the presence of a centrally-located domain termed TAFH. This domain adopts an helical fold with a large hydrophobic groove that forms a binding surface for transcriptional activators such as SP1, NFAT, HP1 and NF-E2 (Amrolia et al. 1997; Kim et al. 2001; Vassallo and Tanese 2002; Hibino et al. 2016; Hibino et al. 2017). Notably, several lines of evidence have shown that TAF4 isoforms with structural TAFH variations are essential for the establishment of differentiation programs for various cell lineages (Kazantseva et al. 2013; Alpern et al. 2014; Kazantseva et al. 2014; Langer et al. 2016; Kazantseva et al. 2016), suggesting a central role for TAF4 as a mediator between gene-specific activators and transcriptional output.

**TAF5**

TAF5 is another of the core TFIID subunits present in two copies and essential for complex assembly and integrity. The C-terminal region of TAF5 contains a highly conserved WD40 repeat motifs; short sequences of 40 aminoacids often terminated in tryptophan-aspartatic acid (WD) (Malkowska et al. 2013; Bieniossek et al. 2013). The WD40 domain of TAF5 adopts a B-barrel structure that requires the TRiC/CCT chaperone to properly fold and is crucial for complex stability (Antonova et al. 2018). Subsequent release from the chaperone requires binding to HFD-containing TAF6/9 heterodimer, leading to the formation of a 3TAFs subcomplex. This 3TAFs complex is the precursor of the proposed nuclear 5TAFs core TFIID particle, which also includes the HFD-containing TAF4/TAF12 heterodimer (see figure 1-9). Because other WD40-containing proteins have been show to bind histones, it has been proposed that TAF5 maintains complex integrity by acting as a
scaffold that stabilizes the binding of HFD-containing TAFs (Tao et al. 1997). Consistent with this hypothesis, a decrease in the levels of TAF5 in Drosophila cells led to a reduction in the levels of TAF6, TAF9 and TAF4 (Wright et al. 2006). In addition, temperature-sensitive mutations in the WD-40 repeats of yeast TAF5 were found to cause broad transcription defects, and their analysis emphasized the importance of these motifs in maintaining the integrity of TFIID (Durso et al. 2001).

The N-terminal region of TAF5 includes two domains termed NTD1 and NTD2 or LisH domains. These two domains are believed to enable the formation of a TAF5 dimer, which will then serve as the above mentioned binding platform for other TAFs. Notably, while various studies have confirmed the presence of two copies of TAF5 within TFIID (Leurent et al. 2004; Bieniossek et al. 2013), recent cryo-EM studies with higher resolution have questioned the existence of an homodimer in vivo, since within the structure of TFIID the TAF5 NTD regions only interact during a transient structural conformation of the complex (Louder et al. 2016; Patel et al. 2018).

**TAF6**

TAF6 is present in two copies, is also part of the cTFIID and another of the HFD-containing TAFs. The TAF6 HFD is located N-terminally, is structurally similar to the fold of histone 4 (H4), and forms a stable heterodimer with the H3-like HFD present in TAF9, another member of the cTFIID (Bieniossek et al. 2013). Early biochemical experiments showed that TAF6/9 was important for binding to the DPE element (Burke and Kadonaga 1997), but this association have been challenged recently by cryo-EM studies showing that TFIID binding to the DPE motif is mediated by TAF1/7 and TAF2 (Louder et al. 2016).

The C-terminal of TAF6 contains several HEAT domain repetas that mediate interactions with TAF9 and with the WD40 domain of TAF5. Besides, within promoter-bound TFIID, the HEAT repeats from the two copies of TAF6 form a homodimer that interacts with TAF8 and TAF2 and structurally bridges the TFIID subunits located at the
upstream and downstream core promoter region (Louder et al. 2016; Patel et al. 2018).

**TAF7**

TAF7 is present in one copy and is the binding partner of TAF1, with who it strongly interacts through a triple barrel located in the N-terminal part of the protein (H. Wang et al. 2014). TAF7 has also been shown to interact with various transcriptional activators (YY1, SP1, c-Jun, USF) as well as PIC components (TFIIH) and elongation factors (Brd4, P-TEFb) (Chiang and Roeder 1995; Munz et al. 2003; Gegonne et al. 2008; Gegonne et al. 2013). These studies have led to the proposition that TAF7 is a key controller of transcription, serving as regulatory checkpoint that mediates the transition from initiation to pausing and elongation. However, none of these functions are well established. More recently, the cryo-EM promoter-bound TFIID structures have shown that together with TAF1, TAF7 is in charge of interacting with the downstream core promoter region, in particular the DPE element (Louder et al. 2016).

Similar to other TAFs, a TAF7 paralog, TAF7L, has been implicated in gametogenesis (Pointud et al. 2003; Cheng et al. 2007; Zhou et al. 2013). During early spermatogenesis, TAF7L accumulates in the cytoplasm and is then imported into nucleus at later stages of the process, an event that coincides with downregulation of canonical TAF7 and upregulation of TBP. The importance of TAF7L for gametogenesis is also evidenced by gene polymorphisms leading to male infertility and germ-cell defects. Notably, whether TAF7L integrates into holo-TFIID in a similar manner to TAF7 remains unclear.

**TAF8**

TAF8 is present in one copy, is another of the HFD-containing TAFs and plays several key roles within TFIID. It contains an N-terminal located HFD that is used for interaction with TAF10, whereas its C-terminal region features a proline-rich fragment essential for TAF2 binding (Trowitzsch et al. 2015). Together, these TAFs form a heterotrimeric complex whose incorporation into cTFIID is thought to drive structural
rearrangements required for the formation of the holo-complex and association with TBP. Importantly, TAF8 also contains a nuclear localization signal (NLS) that is absent in TAF10 and TAF2 and is essential for their import into the nucleus.

As mentioned above, the recent cryo-EM studies showed TAF8 as a key structural component that bridges different lobes of the promoter-bound TFIID (Cianfrocco et al. 2013; Patel et al. 2018). Based on this observation, it would be reasonable to think that upon disruption of TAF8, the integrity of the DNA-bound TFIID complex as well as Pol II transcription would be broadly and severally impaired. Interestingly, while depletion of TAF8 is lethal in mouse embryonic stem cells (mESCs), a patient with intellectual disability harboring a homozygous TAF8 mutation and undetectable TAF8 protein, had preserved RNA polymerase II transcription (El-Saafin et al. 2018). These results suggest that partial TAF complexes, and/or an altered TFIID containing a mutated TAF8, could support human development and are therefore sufficient to assemble a functional PIC.

**TAF9**

TAF9 is present in two copies and is another of the cTFIID subunits, where it forms a stable heterodimer with TAF6 through its N-terminally located HFD. As described previously, the TAF6/9 heterodimer interacts with TAF5 in the cytoplasm and is required for its release from the TRiC/CCT chaperone (Antonova et al. 2018). Moreover, TAF9 has also been involved in recognition of DPE element (Burke and Kadonaga 1997).

The C-terminal region of TAF9 is relatively undercharacterized and is not required for incorporation into TFIID or TAF6 interaction, but its high conservation suggest important roles for gene transcription (Saint et al. 2014). In yeast, mutations in the C-terminal region of TAF9 resulted in a drastic decrease in promoter occupancy of TFIID and the coactivator SAGA.

Notably, a paralog of TAF9, TAF9B, is also able to interact with TAF6 and integrate within TFIID (Frontini et al. 2005). Both proteins are essential for cell viability and thought
to play partially overlapping roles, however, TAF9B has been reported to be particularly important for transcription of HeLa cells and upregulated together with TAF7L and TAF4B during embryogenesis (Frontini et al. 2005; Gura et al. 2020). Moreover, differential induction of TAF9 and TAF9b during apoptosis, together with their different ability to stabilize p53, points to distinct requirements for the two proteins in gene regulation.

**TAF10**

TAF10 is also a HFD member of TFIID and exists in two copies within TFIID. The HFD of TAF10 associates with TAF8 to form the previously mentioned TAF10-TAF8-TAF2 heterotrimer or with TAF3 (Patel et al. 2020). The nuclear import of TAF10 is mediated either by TAF8 or TAF3, both of which contain a NLS. Functionally, several genetic and biochemical studies have shown an essential requirement of TAF10 during early tissue differentiation, but not in terminally differentiated adult cells, suggesting a role establishing specific gene-regulatory programs during development rather than in maintaining complex integrity (Mohan et al. 2003; Indra et al. 2005; Gura et al. 2020). In agreement, TAF10 has been shown to directly interact with the transcriptional activator GATA1 as well as CNOT3, CBX1 and TRIM28 to regulate gene-specific transcription during mouse erythropoiesis (Papadopoulos et al. 2015).

**TAF11**

TAF11 is one of the smallest TFIID subunits. The only known structural feature of TAF11 is a C-terminal HFD that forms a stable heterodimer with TAF13 (Gupta et al. 2017). Together, these two proteins have been shown to be important regulators of TBP (Shen et al. 2003; Gupta et al. 2017). In yeast, inactivation of TAF11/13 does not affect the recruitment of other TAFs to the chromatin, but it greatly reduces the levels of promoter-bound TBP. Although is clear that TAF11/13 control TBP DNA binding, how this regulation occurs *in vivo* is not clear. Biochemical studies have shown that TAF11/13 can form an heterotrimeric complex together with TBP and compete with the TAF1-TAND for
interaction with the TBP DNA binding surface, while the more recent cryo-EM studies suggest that this heterodimer covers the TBP convex surface to avoid premature interaction with TFIIA and TFIIIB (Gupta et al. 2017; Patel et al. 2018; Chen et al. 2021). Finally, it has been proposed that TAF11 entails TFIID with promoter-bound TFIIA.

**TAF12**

As described previously, TAF12 is the heterodimeric partner of TAF4. Like TAF4, TAF12 is present in two copies, contains a HFD and is part of the cTFIID complex. Interestingly, while in humans the best characterized activator binding motifs within TFIID are in the N-terminal region of TAF4, in yeast these are found in the N-terminal region of TAF12 (Patel et al. 2020). In fact, TAF12 has been shown to directly interact with the ATF7 factor and it has been implicated in the progression of acute myeloid leukemia (AML) through association with the oncogenic factor MYB protein (Hamard et al. 2005; Bell and Dawson 2018; Xu et al. 2018).

**TAF13**

TAF13 is present in one copy and is the smallest member of TFIID. Most of its primary sequences occupied by the HFD that uses to pair with TAF11. As described above the TAF11/13 forms a complex with TBP in the context of promoter-free TFIID and regulates TBP availability to interact with DNA (Shen et al. 2003; Gupta et al. 2017).

### 1.9 Assembly of the holo-TFIID complex

One of the most remarkable architectural features of TFIID is the presence of HFDs in ten of the TAFs subunits. These domains were initially found through sequence homology with histone proteins (Kokubo et al. 1994; Hoffmann et al. 1996) and their presence was later confirmed *in vitro* by X-ray diffraction experiments, showing that TAF6-9 can form a tetramer organized by the HFDs (Xie et al. 1996). Following studies showed that the HFDs were responsible for the formation of five different heterodimers comprising TAF3-10,
TAF8-TAF10, TAF11-TAF13, TAF6-9 and TAF4-12 (Birck et al. 1998; Gangloff et al. 2000; Reese et al. 2000; Sanders and Weil 2000; Gangloff et al. 2001). The fact that TAF10 could form a heterodimer with two different partners and that TAF6-9 was found as tetramer in vitro, led to the suspicion that several TAFs might appear in more than one copy within TFIID. The first evidence came from experiments in yeast, uncovering a stoichiometry where six TAFs (TAF4, TAF5, TAF6, TAF9, TAF10 and TAF12) were present in two copies, whereas TBP and the remaining seven (TAF1, TAF2, TAF3, TAF7, TAF8, TAF11 and TAF13) existed in a single copy (Sanders et al. 2002). Later, immunolabeling experiments of TAF5 followed by negative-stain EM showed that this subunit was found in two separate locations within human TFIID (Leurent et al. 2004), further supporting the idea that several TAFs exists in two copies. More recently, several cryo-EM studies in yeast and human TFIID have confirmed these results (Patel et al. 2018; Kolesnikova et al. 2018).

Another interesting finding about the structural organization of the TFIID complex was made by Robert Tjan in 2006. Using Drosophila cells, they showed that depletion of any of the double-copied TAFs (except TAF10), compromised the stability of other TAFs, suggesting that a structural core-TFIID subcomplex exists in vivo (Wright et al. 2006). The prevalence of discrete TAFs heterodimers and the existence of a core-TFIID module led to hypothesize a stepwise model for TFIID assembly. Such a model would start with a core-TFIID and progressive incorporation of the remaining TAF would lead to the formation of the functional holo-complex.

Combinatorial assembly assays using HeLa extracts supported this idea by revealing that TAF8-10 heterodimer could only integrate into TFIID when all core TAFs were present (Demény et al. 2007). This was later demonstrated in vitro using recombinant expression of the core-TFIID and TAF8-10 modules, followed by single-particle cryo-EM analysis (Bieniossek et al. 2013). The reported structures revealed a two-fold symmetric core-TFIID with an interlaced organization and protrusions, whose evenness was broken upon addition.
of the TAF8-TAF10 subcomplex. Presumably, the resulting asymmetry exposes new binding surfaces within TFIID that might facilitate the recruitment of TBP and the remaining TAFs.

The presence of cytoplasmic TFIID subcomplexes also supports a modular assembly. Stable TAF2-8-10 and TAF5-6-9 heterotrimeric complexes, as well as TAF11-13 heterodimers have been found in the cytoplasm of human and yeast, respectively (Soutoglou et al. 2005; Trowitzsch et al. 2015; Gupta et al. 2017; Antonova et al. 2018). TAF1-7 and TAF1-TBP subcomplexes can also form in vitro and co-translation of TAF6-9, TAF8-10 and TBP-TAF1 has been shown in vivo. It is important to note that since only a subset of TAFs possess a NLS, the import of TAF complexes into the nucleus greatly depends on the availability of their respective interaction partners. Besides, these data also suggest that TFIID assembly is a dynamic process and exchange between different submodules may serve as checkpoints for TFIID assembly and transcription initiation. However, as of today, the in vivo relevance of a stepwise model for TFIID assembly remains to be determined. Figure 1-9 shows a hypothetical model of how different modules of TFIID may be imported into the nucleus to assemble into a functional holo-complex.
Introduction

Figure 1-9. Hypothetical model for TFIID assembly

Several TFIID subcomplexes assemble in the cytoplasm and are imported into the nucleus due to the presence of nuclear localization signal sequences on TAF4, TAF8, TAF1 and TAF3. Assembly of holo-TFIID then likely occurs inside the nucleus. More about the distinct “abc” TFIID lobes is discussed below in section 1.10.

1.10 Structural basis of TFIID promoter recognition

With the aim to better understand the role of TFIID in transcription and its mechanism for promoter recognition, tremendous efforts have been made to characterize its DNA-bound structure. While traditional X-ray crystallography experiments were limited due to poor complex stability, substoichiometric composition and inherent flexibility, initial structural descriptions were conducted using low-resolution electron microscopy (EM). These reported a saddle-like structure with a high degree of plasticity and three defined lobes named A, B and C (Andel et al. 1999; Brand et al. 1999; Grob et al. 2006; Cler et al. 2009). Further studies showed that upon DNA binding, TFIID undergoes a series of step-wise conformational changes that shift the complex from a canonical to an extended state (Figure 1-10A) (Cianfrocco et al. 2013). The major difference between those states lies in the location of the lobe A, which travels to opposite sides of the complex while switching
contacts from lobe C (canonical/inhibited) and B (extended/engaged), presumably to finely position TBP at the TATA-box.

Although these findings set a framework for understanding the mechanism of promoter recognition, detailed characterization of the holo-TFIID as well as the exact location of each subunit within the structure was still limited due to the intrinsic flexibility of the complex and the low resolution of the method. The breakthrough came quite recently, thanks to the rapid advances in the cryo-EM technology and by the improvement of the crosslinking conditions during TFIID sample preparations, which have allowed to model the promoter-bound and in solution structure of the TFIID complex at high resolution (Louder et al. 2016; Patel et al. 2018). The BC module was more rigid compared to A, allowing its structural characterization at ~ 4 angstrom (Figure 1-10B). Lobe C includes a TAF1-7 dimer, TAF2 and two HEAT-repeat domains from a TAF6 homodimer, whereas Lobe B is composed of three pairs of TAF-HFDs, TAF4-12, TAF6-9 and TAF10-8, arranged by interactions with a single TAF5 WD40 domain. Lobe B and C are connected by the TAF8-HFD, which extends to the TAF6 homodimer and anchors both proteins to TAF2, an interaction that is in agreement with the proposed role of TAF10-8 in driving TAF2 incorporation into a core-TFIID.

Lobe A was harder to characterize, but using focused refinement and based on the subunit stoichiometry of Lobes B and C, an architectural model of ~10 angstroms was constructed (Figure 1-10C). The refined model suggested an octamer of HFD-containing TAFs organized by the TAF5 WD40 domain that resembles lobe B. However, apart from TAF6-TAF9 and TAF4-TAF12, this lobe includes a TAF11-13 heterodimer and TAF10 partners with TAF3 instead of TAF8. Besides, decorating the top of this nucleosome-like structure is TBP, which anchors to this lobe through interactions with the TAF11-13 HFDs.

In the canonical state, lobes A and C are assumed to be connected through the TAF1 N-terminal domain (TAND), which binds the concave surface of TBP and together with
TAF11/13 prevent its interaction with DNA \((\text{Figure 1-10D})\). Upon TFIID DNA binding, TBP relocates upstream of the TSS accompanied by the TAFs in the lobe A and scans for the TATA-box. Binding to the TATA-box releases TBP from its TAF1-TAND-inhibited state and is aided by TFIIA, whose binding to the promoter is further stabilized by interactions with the lobe B-localized TAF4. Besides, this causes steric clashes with TAF11-13, which detaches from the TBP convex surface and opens space for TFIIIB interaction, which binds opposite to TFIIA and allows progression of the PIC assembly by the subsequent recruitment of TFIIF-Pol II, TFIIE and TFIIH.

It is interesting to note that despite their homology to histones, TAFs containing HFD within lobes A and B do not seem to use these motifs for binding DNA, but rather as a scaffold that allows modular TFIID assembly and complex stability \((\text{Patel et al. 2020})\). In fact, most TAF-promoter contacts occur at Lobe C (which lacks HFDs) via the TAF1 winged-helix (WH) domain and the homology-based aminopeptidase (APD) region of TAF2, although a loop extending from TAF4 in lobe B also contacts DNA right downstream of the TATA-box \((\text{Louder et al. 2016})\). Notably, while previously it was belived that TAF1 and TAF2 bind the core promoter through the Inr \((\text{Chalkley and Verrijzer 1999})\), the current cryo-EM models show that these proteins mainly contact the core promoter region encompassing the MTE and DPE elements with only a part of TAF1 protruding near the TSS \((\text{Louder et al. 2016})\). This was surprising as the MTE/DPE motifs had been long thought to be recognized by the TAF6-9 heterodimer \((\text{Burke and Kadonaga 1997})\), however, in light of the structural data is it reasonable to think that such DNA-binding was the result of (TAF9)-TAF6-TAF2 interaction caused by the linker region of TAF8-HFD.

Overall, the mechanistic insights coming from the human TFIID cryo-EM data challenge our previous intuition about transcription initiation as they imply that binding to the downstream promoter region by TAFs, and not TBP-TATA interactions, is the first step of PIC assembly. Importantly, the similar yeast TFIID structure did not show any re-
arrangements upon binding to an endogenous TATA promoter (Kolesnikova et al. 2018). While this may simply be due to technical differences during sample preparation, it could also mean that an alternative mechanism for TBP-loading occur in yeast. However, it also raises the question of whether lobe A relocation also occurs in the absence of strong Inr and MTE/DPE motifs, which may transiently stabilize the promoter-bound canonical state of TFIID. Besides, the proposed model for TBP loading was obtained using the synthetic super core promoter (SCP) (Juven-Gershon et al. 2006), which gathers strong elements that rarely occur together in vivo, raising the question of how TFIID recruitment and TBP loading occurs at different types of endogenous promoters. Thus, while these recent cryo-EM studies have set the basis for promoter recognition by TFIID, the prevalence of such TBP loading model in vivo and in other organisms remains to be demonstrated.
Figure 1-10. The structure of TFIID and model for DNA binding

(A) Organization of human TFIID into three lobes based on early cryo-EM maps and structural changes upon DNA-binding is shown in the center. Current coordinate models shown next to them for comparison (B) Structure of promoter-bound TFIID and TFIIA in side view (top) and top view (bottom). Protein subunits are shown in colour and labelled; the TBP-induced DNA bend and the location of the transcription start site (Inr) are labeled. (C) Structure of TFIID in solution (canonical state). (D) Before DNA-binding, the lobe A is in contact with lobe C and TBP is found in its canonical or inhibited state via TAF1 and TAF11/13 interactions, which respectively bind to the concave and convex surface of TBP. Upon DNA binding, lobe A separates from lobe C and starts scanning DNA upstream of the TSS until TBP finds the TATA-box. Such interaction releases TBP from TAF1 and TAF11/13 contacts and allow the recruitment of TFIIA first (shown) and TFIIIB after (not shown), allowing progression of the forming PIC. Adapted from (Greber and Nogales 2019).
1.11 Genome-wide mapping of transcriptional complexes at high-resolution can be used to determine their DNA-bound organization

The vast majority of our knowledge about transcription comes from reconstituted *in vitro* systems and structural studies. These approaches are extremely powerful to obtain mechanistic insights since most variables within the system can be finely tuned. However, because these experiments occur in highly controlled conditions, most often they do not recapitulate the native environment within the cell nucleus. For example, the recent TFIID high-resolution structures were obtained using the synthetic SCP, mild-crosslinking conditions and the only other BTF present was TFIIA. Besides, such systems are usually slow to build and limited in the number of genes that can be tested in one experiment. Within this context, the development of Next-generation sequencing (NGS) and the consequent dropping in sequencing costs have represented a unique opportunity to study transcription regulation *in vivo* and genome-wide fashion. As a result, a vast number of sequencing technologies have been developed during the last 10 years with the aim of measuring a variety of transcriptional features, from chromatin accessibility to RNA expression.

One fundamental task of transcriptional studies is to determine which DNA sequences are bound by particular transcription factors across the genome. As of today, chromatin immunoprecipitation coupled to deep sequencing (ChIP-seq) (Raha et al. 2010) is the most commonly used method to detect the DNA-binding location of proteins genome-wide (*Figure 1-11, left*). This technique involves formaldehyde cross-linking of all DNA-protein interactions within the cell, followed by sonication of the chromatin into smaller fragments and later immunoprecipitation of the protein of interest, pulling only the DNA fragments which are bound by the targeted factor. Such DNA fragments are then purified and used for library preparation before submission to high-throughput sequencing.

ChIP-seq is a valuable tool for studying protein-DNA interactions because it is easy to implement and many tools are readily available for its analysis. However, the data has
low resolution and is high in background, which limits the interpretability of the results. This occurs because the length of ChIP-seq DNA fragments (~250 nt) is defined at the sonication step and these are much longer than the usual size of regulatory DNA-binding motifs (~10-15 nt). This is an important downside of the technique as the ultimate purpose behind these experiments is to determine the precise sequence of nucleotides involved in transcription factor binding. To overcome this problem, several approaches have been developed over the past years to improve ChIP-seq resolution.

The most popular is called ChIP-exo (Rhee and Pugh 2012a), which introduces a lambda exonuclease digestion step during library preparation when DNA is still bound to the protein (Figure 1-11, middle). Lambda exonuclease digests protein-free DNA from the 5’ to the 3’ end and stops when it clashes with the border of the protein-DNA crosslink. While this provides precise information about the DNA binding location of transcription factors, ChIP-exo data is overall noisy and often also hard to interpret. To improve the efficiency of the ChIP-exo protocol, our lab more recently developed a variation named ChIP-nexus (He et al. 2015), which replaces intermolecular ligation with a more productive single-molecule circularization during library preparation and introduces randomized barcodes to account for PCR over-amplification (Figure 1-11, right).

Although the primary goal of ChIP-exo and ChIP-nexus is to precisely reveal important regulatory sequences, their footprints can also be informative about the DNA-bound organization of transcriptional complexes. Several studies have shown that precise mapping of the PIC factors with ChIP-exo and ChIP-nexus allows for direct comparison with structural models (Rhee and Pugh 2012b; Shao and Zeitlinger 2017; Yamada et al. 2019). Moreover, these methods can be used in combination with transcription inhibitors to detect fine changes in protein location between different states of the transcription cycle (Shao and Zeitlinger 2017). Remarkably, ChIP-nexus can also be performed in a plasmid and closely recapitulates the endogenous binding, opening the door for the development of
massively parallel reporter assays to test the functionality of bonafide DNA-regulatory regions (Shao et al. 2019). Thus, ChIP-nexus represents a robust method to interrogate protein-DNA binding at an unprecedented resolution and can shed light on transcription regulation when used in combination with other tools like transcriptional drugs or reporter assays.
Introduction

**Figure 1-11. Common methods to detect TF location in vivo**

**ChIP-seq** (left), is the most frequently utilized technique to determine the genome-wide location of transcription factors. However, it lacks the resolution to precisely detect the exact DNA-sequences transcription factors recognize. **ChIP-exo** (middle) implements a lambda exonuclease digestion step, which degrades single-stranded DNA (ssDNA) from 5’ to 3’ until it finds a physical barrier, achieving base-pair resolution by mapping the exonuclease stop bases back to the reference genome. **ChIP-nexus** (right) reduces noise and increases robustness by the addition of adaptors in just one single ligation step and the introduction of a randomized barcode to track PCR-duplicates.

51
Aims of this thesis

The main obstacle to study TFIID binding at endogenous promoters has been the difficulty to obtain high-resolution structures, which can be mainly attributed to the poor stability of the complex, its flexible nature and sub-stoichiometric composition. While recent advances in cryo-EM methods have recently made possible to define the structure of the promoter-bound human TFIID at near atomic resolution, these studies are based on the synthetic Drosophila super core promoter (SCP), which includes elements that rarely occur together in vivo such as the TATA-box and MTE/DPE.

Although that biochemical and structural studies have been fundamental for understanding the basis of TFIID promoter recognition, deeper questions remain to be answered. For example, are TBP and TAFs present at all promoter types? Does the TFIID binding organization change depending on core promoter architecture? Does TFIID remain bound to the core promoter after Pol II initiation? These questions are not easy to address. First, because as explained, TFIID structural studies are challenging and not feasible at large scale and second, because traditional genomic approaches like ChIP-seq usually lack the resolution to detect fine changes in TF binding. Importantly, higher resolution assays to detect the DNA-binding location of transcription like ChIP-exo or ChIP-nexus are available and have been proved to be useful in addressing the spatial organization of transcriptional complexes. However, when applied to TFIID, most studies have targeted just one or two proteins to proxy for the whole complex and little or none focus has been put in addressing changes within TFIID architecture or composition between promoters types.
The aim of this thesis work is to fill this gap by 1) using ChIP-nexus to investigate the genome-wide binding of all TFIID subunits at different types of promoters and 2) investigate whether TFIID is associated with the paused Pol II. The goal was to shed light on how TFIID regulates transcription at distinct promoter types and help us understand how gene expression dynamics are regulated.
Chapter 2 Results

2.1 Holo-TFIID is present at all active promoters

A question that has long remained unclear is whether there are promoter-specific differences in TAF composition or whether all TAFs are present at promoters. To investigate this in vivo, we performed ChIP-nexus experiments on TFIID subunits in Drosophila kc167 cells using ChIP-nexus (Figure 2-1A). This ChIP-exo protocol maps protein occupancy across the genome at near base resolution through digestion with exonuclease, which stops when it encounters a protein crosslinked to DNA. We performed ChIP-nexus using polyclonal antibodies raised against TBP or each of the thirteen TAFs of TFIID. To capture TFIID binding specifically at the PIC stage, we performed the same experiments after treating the cells with the transcription inhibitor triptolide (TRI), which blocks the helicase activity of TFIIH and therefore prevents Pol II transcription initiation (Titov et al. 2011; Bensaude 2011; Q.-L. He et al. 2015).

Mapping the stop bases of each ChIP-nexus experiment yielded high-resolution footprints at promoters for each TFIID subunit and biological replicate experiments were highly reproducible (example in Figure 2-1B). This was also true for TAFs that are not known to bind DNA directly, consistent with formaldehyde frequently producing protein-protein crosslinks in addition to the relatively rare protein-DNA crosslinks, enabling DNA that is indirectly bound to be pulled down in ChIP. As a result, proteins that physically interact with each other in a multi-subunit complex often show similar footprints in ChIP-exo/ChIP-nexus experiments (Rhee and Pugh 2012b; Yamada et al. 2019). Thus, although the exact protein-protein interactions cannot be inferred, we expect that the ChIP-nexus footprints of the TFIID subunits are a reflection of the structure of TFIID in vivo.
We then calculated the average ChIP-nexus footprints for each TFIID subunit. To preserve the high resolution of the ChIP-nexus data, we first derived precise TSS annotations for our cell type using CAGE-seq data (Shiraki et al. 2003), selected active and then averaged each factor’s data (-200 to 200 bp from TSS) across the 1000 promoters with highest TAF1 occupancy (Figure 2-1C). Our results show that the 14 TFIID subunits make DNA contacts at roughly six distinct positions upstream and downstream of the TSS (-150/-50, -30, -18/-14, +10, +19 and +32) and that each subunit has preferred interactions. Based on the location of these interactions, the TFIID subunits can be grouped into four modules (upstream broad, upstream narrow, middle, and downstream) (Figure 2-1D), a classification that was confirmed by PCA analysis (not shown).

Notably, the four modules and their preferred DNA contacts upstream and downstream of the TSS are remarkably consistent with cryo-EM structures of a promoter-bound human and yeast TFIID complex (Figure 2-1E). Thus, the results support the hypothesis that many DNA contacts of TFIID subunits occur indirectly through other subunits of the complex and are indicative of the structural organization of the complex.

The broad upstream module is composed of TAF4, TAF10, and TAF12 and is characterized by fuzzy footprints upstream of the TSS, which appear as broad and spread out signal from ~ -150 to -50 bp in the average profiles (Figure 2-1D, region 1). This DNA region often functions as a binding platform for TFss, especially at housekeeping genes whose regulatory sequences lie within and immediately upstream of the core promoter (Zabidi et al. 2015). These broad footprints could therefore be due to protein-protein interaction with TFs, which TAF4, TAF12 and TAF10 are known to engage (Hamard et al. 2005; Wei et al. 2007; Papadopoulos et al. 2015; Kazantseva et al. 2014). Since TAF4/10/12 are present in two copies in TFIID, these interactions could occur through the copies in the upstream module (Figure 2-1E) or through the copies in the flexible lobe (Patel et al. 2018).
The narrow upstream module, consisting of TAF5, TAF9, TAF13 and TBP, is the module most closely associated with the PIC. It shows altogether three distinct footprints centered at -30, -18 and -14 bp (Figure 2-1D, regions 2 and 3). Unexpectedly, the main footprint of TBP was found at the PIC site (-18 and -14 bp) and not near the TATA-box at the -30 bp position. Instead, the -30 bp position had a footprint by TAF13, which directly interacts with TBP to position it upstream to the TATA box (Shen et al. 2003; Gupta et al. 2017). This suggests that TFIID undergoes minor rearrangements during PIC formation, consistent with structural studies suggesting that TBP loses direct contact with the rest of TFIID in order to allow for new interactions with other PIC components (Patel et al. 2018). Furthermore, weaker footprints at the site of the PIC were observed for all TAFs, suggesting close contacts during PIC assembly.

The middle module, consisting of TAF6, TAF8 and TAF11, shows footprints both upstream and downstream of the TSS, suggesting it bridges the upstream and downstream lobe of the complex. Indeed, TAF6 and TAF8 are found exactly at that middle location in the cryo-EM structures (Figure 2-1E). The inclusion of TAF11 in the module was however unexpected, since TAF11 forms a stable heterodimer with TAF13 (Gupta et al. 2017), which was classified as part of the upstream module. The simplest explanation is that TAF11, but not TAF13, makes direct contact with TAF6 and TAF8 after releasing TBP. This is plausible since components of the flexible lobe, to which TAF11/13 belong but which is not fully resolved, are relocated when TFIID is fully engaged with the promoter (Patel et al. 2018). This association is also consistent with the yeast TFIID structure where the TAF11/13 heterodimer is adjacent to TAF6/9 (Kolesnikova et al. 2018) (Figure 2-1E).
Finally, TAF1/7, TAF2 and TAF3 were part of the downstream module and showed three clustered footprints at +10, +19 and +32, suggesting simultaneous engagement with the MTE and DPE elements in the downstream core promoter region (Figure 2-1D). This is consistent with the human cryo-EM structure where TAF1/2/7 form a complex that contacts the DPE (Louder et al. 2016), but not with previous biochemical experiments suggesting that the DPE is recognized by TAF6/9 (Burke and Kadonaga 1997). The addition of TAF3 is interesting since just like TAF11/13, TAF3 originates from the flexible lobe and its position in promoter-bound TFIID has not been clear (Figure 2-1E). However, the association of TAF3 with TAF1/2/7 is supported by a recent cross-linking mass spectrometry analysis in yeast (Kolesnikova et al. 2018) and the footprints in the downstream promoter region are consistent with TAF3’s ability to interact with the H3K4me3 mark through its PHD domain (Vermeulen et al. 2007; Lauberth et al. 2013), which is typically present at the +1 nucleosome located downstream of TFIID.

Having confirmed the specific binding of individual TFIID subunits, we next asked whether the subunit composition might vary between promoters. If some subunits are present in different amounts, we would expect the amount of signal across promoters to be more variable, especially as compared to other basal transcription factors such as TFIIA, TFIIB and TFIIF. Furthermore, we expect that TFIID subunits that tightly interact with each other to be more correlated with each other than non-interacting subunits. We therefore calculated for each subunit the total signal across each promoter (from -50 to +50 bp) and determined the Pearson correlation coefficient (PCC) between all pairs of TFIID subunits after TRI treatment, as well as Pol II and other basal transcription factors (TFIIA, TFIIB and TFIIF) (Figure 2-1F).
The results show that the levels of TFIID subunits are overall highly correlated with each other and also correlate well with TFIIA, Pol II, TFIIB and TFIIF (median PCC = 0.69). The correlations with the basal transcription factors are only slightly lower without TRI treatment (median PCC = 0.65), consistent with the higher signal (and reduced noise) after TRI treatment and TFIID’s main function in transcriptional initiation (Figure 2-1E). These results suggest that all subunits of TFIID are present at active promoters and that the levels are roughly proportional to the amount of other PIC components.

Despite the overall high Pearson correlations, hierarchical clustering suggested that there were differences that were biologically meaningful based on previous knowledge (Figure 2-1F). The TFIID subunits that were most highly correlated and clustered first were TAF4/12, TAF1/2/7 and TAF6/9 (Figure 2-1F left), all of which are known to form strong protein-protein interactions (Figure 2-1F right). Further clustering revealed that TAF4/12 occupancy levels most closely resembled those of other subunits from the upstream modules, while those of TAF2/7 clustered with TAF1 and the middle module subunits. Notably, TAF3 clustered with the upstream modules. Since TAF3 is part of the flexible module, this could suggest that the flexible lobe interacts with upstream TFs at some promoters.

Lastly, TBP clustered more closely with TFIIA, other basal transcription factors and Pol II than with other TFIID subunits (Figure 2-1F). This is again in agreement with the close association of TBP with TFIIA and with TBP being a more mobile component of TFIID that is rate-limiting for transcription initiation (Patel et al. 2018).
Figure 2-1. DNA-bound TFIID is organized into three major modules

(A) Schematic of the experimental workflow followed in this work. (B) ChIP-nexus data in Drosophila Kc167 cells show distinct binding location at individual promoters (e.g., FASN1). (C) TFIID occupancy at active promoters (n=1000) sorted based on decreasing TAF1 signal shows strong correlation between all subunits. (D) Average binding profile of all TFIID subunits at top 1000 promoters with the highest TAF1 signal. Subunits were manually overlapped based on the similarity in their binding profiles and therefore the relative differences in occupancy seen for overlapping TAFs is completely arbitrary. (E) Cryo-EM structure of the promoter-bound human TFIID complex
showing the main areas of contact captured by ChIP-nexus. (F) Pair-wise Pearson correlation between TFIID subunits and PIC members calculated using the total ChIP-nexus signal in 101-bp window centered at the TSS. Hierarchical clustering (ward.D method) was performed using the Pearson correlation coefficients (PCC) and was consistent with several known protein-protein interactions. Structural data was obtained from RCSB Protein Data Bank (PDB) under accession numbers 1NVP (TFIIA-TBP), 4OY2 (TAF1-2-7), 1H3O (TAF4-12), 6MZM (TAF6-9).
2.2 TFIID organization at focused vs dispersed promoters

Core promoters are often classified based on their transcription initiation pattern, which range from focused/sharp to broad/dispersed (Hoskins et al., 2011; Ni et al., 2010). By definition, focused transcription occurs when there are one or several TSSs that are clustered close to each other in a narrow region. In contrast, dispersed or broad initiation occurs when there are multiple weak or strong TSSs distributed over a wider region, usually 50-100 nt (Vo Ngoc, Wang, et al. 2017). While the spectrum of initiation patterns and their associated DNA-motifs and functionalities are well characterized, it is unclear whether TFIID (and the PIC) binding differs between those promoters.

To examine the correlation between TFIID binding and transcription initiation pattern genome-wide, I used CAGE-seq data to form TSS clusters (TCs) and separated focused from dispersed promoters using the calculated width of such clusters (Figure 2-4A-B), defined as the region where 80% of transcripts initiate (Haberle et al. 2014; Lu et al. 2020). Consistent with previous studies, our data showed that focused TCs (width <= 9) were enriched in motifs like TATA, Inr, MTE, DPE, and PB (Figure 2-4C), usually associated with tissue-specific functions (Zabidi et al. 2015; Haberle et al. 2019). On the contrary, dispersed TCs (width > 9) were enriched in motifs like DRE, Ohler1, Ohler6 and Ohler7 (Figure 2-4C), more often present at ubiquitously-expressed housekeeping genes (Zabidi et al. 2015; Haberle et al. 2019). An exception to this classification is the TCT element, which encompasses transcription of all ribosomal protein genes but presents focused initiation despite being expressed in all cell types and all stages of development (Vo Ngoc et al. 2019).
Due to the lack of a well-defined center of alignment, I reasoned that ChIP-nexus profiles would likely be noisier at dispersed promoters. To test this, I compared the binding of TAF2, which contacts DNA directly, and other basal factors (TFIIB and Pol II) between the focused kik1 and the dispersed Gβ13F promoters. As expected, I observed dramatic differences in their binding profile and occupancy (Figure 2-4D). At kik1, TAF2 and basal factor footprints were sharp and located at well-defined positions relative to the TSS. In contrast, at PŻg, binding footprints were spread and fuzzy, consistent with the idea that the transcription machinery can assemble at distinct locations over a broad region at dispersed promoters. In agreement, the average binding profile of TAF2 genome-wide was sharper at focused than at dispersed promoters (Figure 2-4E).

Besides, the total signal of TAF2, TFIIB and Pol II was significantly lower at dispersed TCs (Figure 2-4F). This could be caused either because of low expression or unstable DNA-binding at these promoters. To discriminate between these possibilities, I compared the RNA expression of focused and dispersed promoters using RNA-seq and CAGE-seq data (Figure 2-4G). Because RNA levels were comparable in both groups, it is reasonable to conclude that TFIID and the whole PIC is overall less stable at dispersed/housekeeping promoters.
Figure 2-2. TFIID is more stable at focused/regulated promoters

(A) Schematic workflow of a CAGE-seq experiment. (B) CAGE signal (RPM) at active promoters sorted by the width of TSS clusters (TCs) separates focused from dispersed promoters. (C) Focused and dispersed promoters enriched for distinct types of core promoter elements (Fisher’s exact test with multiple testing correction, p-value < 0.05). (D) A single gene example showing that regardless of having similar levels of initiation, ChIP-nexus footprints look sharp at focused promoters compared to the dispersed. (E) The average footprints of TAF2 look sharp and with well-defined positions around the TSS at the focused promoters, but not at dispersed promoters. (F) Comparing the ChIP-nexus occupancy of the TAF2 and TFIIB at focused vs dispersed promoters shows that the PIC is more stable at focused promoters despite having overall less expression levels (Wilcoxon test). To account for the broad signal at dispersed promoters, total occupancy was calculated in 301-bp window centered at the TSS.
2.3 The binding profile of TBP, but not TAFs, is promoter-specific

The good correlation between TFIID components and other basal transcription factors suggested that TFIID is found at all promoters regardless of promoter type, but did not rule out differences in the binding footprints or total occupancy between promoter types. To test this, we classified promoter types based on the presence of specific core promoter elements located at a defined position with respect to the TSS as defined by the CAGE-seq data (Ohler et al. 2002; FitzGerald et al. 2006; Vo Ngoc et al. 2019) (Figure 2-5A and Table 4-2), and examined the binding profile and total occupancy of the TFIID (Here, TBP and TAF2 are used as examples).

Our analysis revealed that the average binding profile of TBP was notably different between promoter types (Figure 2-2A). TBP showed footprints upstream (~ -18 bp) and downstream (~ +30) of the TSS on all promoter types, consistent with our previous observations (Shao and Zeitlinger). At TATA promoters, however, TBP showed a much sharper and stronger footprint upstream at -18 bp. This suggests that the presence of TATA at ~-30 bp, to which TBP binds with high affinity *in vitro*, is indeed associated with this upstream footprint of TBP. The upstream footprint was not lower and more spread out at the other promoter types because the total TBP occupancy was particularly high at TATA promoters (Figure 2-2C) and this was due to a high ratio between upstream and downstream TBP (Figure 2-2D). The preferred upstream binding at TATA promoters was not observed to the same extent for TAFs (e.g. TAF2 in Figure 2D), although some TAFs showed a bias due to their broad upstream footprints.
In contrast to TBP, the remaining TFIID components were highest at the DPR promoters (Figure 2-2B,C). Although this is consistent with observations that TATA-less promoters are more dependent on TAFs, we noted that TATA promoters still had strong and sharp footprints of all TFIID components, suggesting that TFIID assembles at TATA promoters in a manner similar to other promoters. Only housekeeping genes showed substantially lower occupancy of all TFIID components. This was not due to the dispersed nature of Pol II initiation seen for these promoters or due to lower expression. Furthermore, TFIIA, TFIIB and TFIIF were also present at lower levels, despite the high expression of these genes. This could indicate that the PIC is overall less stable at housekeeping promoters.

Our results so far suggest that TAFs are found at all promoter types with similar patterns, while TBP shows promoter-specific profiles. We therefore tested whether this pattern was specific enough to classify promoters into promoter types based on their individual base-resolution TBP binding profile using K-means. We used active promoters with focused initiation (n=1078) and used a gap-statistic method to determine the optimal number of clusters, which was four (see methods). This simple approach identified TATA promoters, DPR promoters and two types of HK promoters based on their underlying sequence patterns and motif contents (Figure 2-3A-E). Consistent with our previous analysis, upstream TBP was higher at promoters enriched in TATA (cluster 1), while at promoters enriched in sequences like Inr, MTE or DPE/PB (cluster 4), most of TBP was located downstream of the TSS. Housekeeping motifs were enriched in clusters 2 and 3, which had overall lower TBP occupancy. Notably, the expression levels observed across clusters (Figure 2-3F) suggest that the specific differences in binding observed between promoter types are not due to transcriptional activity.
Figure 2-3. The binding profile of TBP is promoter specific

(A) Different promoter types selected based on the presence of known *Drosophila* core promoter elements. (B) TBP and TAF2 ChIP-nexus binding profiles at the different types of promoters. Beyond the differences observed in occupancy, the binding profile of TAF2 is consistent across promoter types, while TBP seems promoter specific. (C) Total signal distribution of TBP and TAF2 promoter types. TAF2 occupancy is lower at TATA with respect to the DPR promoters. (D) The proportion of TBP binding upstream and downstream of the TSS changes between TATA and -DPR promoters, while for TAF2 remains the same. A Wilcoxon test was used for all the statistical comparisons.
Figure 2-4. Clustering genes based on TBP binding profile distinguishes between promoter types

(A-B) TBP binding patterns at K clusters. (C) Distribution of the total TBP signal upstream and downstream of the TSS at the selected clusters (Wilcoxon test) (D) DNA-sequence content heatmaps of the promoters contained in the different clusters. (E) Expression levels distribution of promoters contained at the four clusters measured by RNA-seq in TPM and CAGE-seq in RPM.
2.4 TBP appears to bind downstream of the TSS before PIC assembly

The overall structure of TFIID appears to be similar between promoter types, yet the TBP binding profile is remarkably promoter-specific. Together with previous studies, this suggests that the dynamics of TBP differs between promoter types. TBP has many binding partners and is handed over between them, including TAF1, TAF11/13, TFIIA and NC2/BTAF1. Most notably, cryo-EM data suggested that the flexible lobe relocates TBP from the downstream lobe (the canonical state) to the upstream lobe (rearranged state) during promoter scanning and engagement (Patel et al. 2018). This movement delivers TBP to the DNA location where the PIC forms, but whether the canonical state in which TBP is associated with the downstream lobe also exists in a promoter-bound form is not known. Since we observed a downstream binding footprint of TBP, it is tempting to speculate that this is indeed the promoter-bound scanning state of TFIID, which occurs before TBP is loaded to the upstream location. However, there are alternative explanations for the downstream footprint of TBP and, thus, we first set out to test alternative models.

The simplest explanation is that TBP is not found downstream and that the downstream footprint arises from indirect DNA-binding through the downstream TAFs during PIC assembly. We reasoned that if this were the case, then these indirect downstream contacts should be increased after triptolide treatment, which puts TFIID preferentially into a PIC-like state. However, our data showed a significant decrease in downstream TBP at both TATA and DPR promoters and caused a switch in the upstream vs downstream binding ratio at TATA-less promoters (Figure 2-4A and B, top). By contrast, the TAFs of the downstream module (e.g. TAF2) showed a drastic increase in binding, and the proportion of signal upstream and downstream of the TSS remained the same at both promoter types and in both conditions (Figure 2-4A and B, middle). These data argue that the downstream TBP
footprint does not occur during PIC assembly and thus likely arises from a different state of TFIID.

Another possibility is that the downstream TBP is associated with paused Pol II. Although TBP has not been functionally implicated in Pol II pausing, we considered this model because we have previously observed that under normal control conditions, TBP is preferentially found at stably paused promoters and its footprint matches the position of the paused Pol II (Shao and Zeitlinger 2017). However, whether both binding events occur at the same time is not known. Conveniently, TRI treatment also pinpoints to promoters with stable Pol II pausing. Since TRI treatment prevents new Pol II initiation, only DPR promoters with stable Pol II pausing remain paused after 1 h of TRI treatment, while TATA promoters with short Pol II half-lives have lost all Pol II pausing (Shao and Zeitlinger 2017; Shao et al. 2019) (Figure 2-4A bottom). We reasoned that if TBP was directly or indirectly associated with paused Pol II, the DPR promoters should have higher TBP downstream occupancy than TATA promoters after TRI treatment, but this was not the case (Figure 2A, B). Furthermore, we expected that DPR promoters would still show a preference towards downstream versus upstream TBP as observed under control conditions. This was also not the case (Figure 2-4A, B). These data therefore suggest that TBP is found at the downstream location neither during PIC formation nor during Pol II pausing.

We therefore hypothesized that TBP is bound downstream before it is loaded upstream for transcription initiation, thus before TFIIA and other PIC components bind. These results are consistent with our previous finding that TBP clustered closer to other basal factors than with TFIID.

An important aspect to consider for this model is that if the downstream pre-loading of TBP truly occurs before the PIC forms, then other BTFs should not be present at the TATA-less promoters in steady-state conditions. To test this, we compared the binding occupancy of the early PIC components TFIIA, TFIIB and TFIIF between TATA and
TATA-less promoters (Figure 2-4C and D). As expected, very low levels of these factors were found at TATA-less, but not at TATA-containing promoters, indicating that the PIC binding is more stable when a TATA-box is present. However, a significant increase in the occupancy of TFIIA, TFIIB and TFIIF was observed after triptolide treatment at both, TATA and TATA-less promoters, which is consistent with the switch in upstream vs downstream binding observed for TBP after blocking transcription initiation.

Taken together, our data suggest that regardless of core promoter sequence, TBP is first loaded downstream of the TSS and only after it has been relocated upstream other PIC components can be recruited (Figure 2-4E). However, the differences in TBP binding under steady-state conditions are likely to reflect differences in the transcription dynamics between TATA and TATA-less promoters. We discussed the implications of this model in the section 3.3.
Figure 2-5. TBP binds downstream of the TSS before PIC assembly

(A) Comparison of the binding profile and total upstream and downstream occupancy of TBP, TAF2 and Pol II after 1 hour of triptolide treatment shows that the downstream binding of TBP does not correlate with PIC assembly or Pol II pausing (Wilcoxon test). (C-D) Consistent with TBP relocation model, strong occupancy of basal transcription factors at TATA-less promoters is only seen after PIC assembly is blocked by triptolide treatment. (E) Model depicting the TFIID binding states at TATA and TATA-less promoters during steady-state conditions and during PIC assembly.
Chapter 2

2.5 Analysis of TBP and Pol II ChIP-nexus data at dual promoters during early Drosophila embryogenesis reveals no association between TBP and pausing

In the previous analysis I showed that there seems to be no apparent association between TBP and the paused Pol II. While the presence of the downstream footprint of TBP correlates with promoter elements that also influence pausing, the data suggested that these two binding events occur at different transcriptional stages. I concluded this because after triptolide treatment, Pol II binding could still be seen downstream of the TSS at a proportion of TATA-less promoters due to stable pausing, but this did not occur for TBP. However, the overall levels of paused Pol II were also significantly decreased between triptolide and control conditions, and one could argue that the downstream footprints of TBP may be in fact associated with Pol II pausing, but not tightly enough so that a redoubt of binding could be seen even after triptolide, as occurs with Pol II. Thus, I took an orthogonal approach to further investigate the relationship between the downstream binding of TBP and the Pol II pausing.

In Drosophila, there is a small group of promoters known as “dual” that undergo a switch from a non-paused to a paused state during the mid-blastula transition (MBT) (K. Chen et al. 2013), a pivotal event that occurs early in development. Based on this, I reasoned that these promoters could be used as a system to test whether the binding profile of TBP “mimics” the one of Pol II as pausing becomes established. Therefore, I set out to examine the binding of TBP and Pol II at the dual promoters during the transition from pre-MBT to post-MBT. Drosophila embryos were collected by traditional means at 0-2 (pre-MBT), 2-3 (early-MBT), 3-4 (late-MBT) and 4-8 (post-MBT) hours after egg deposition. To avoid contamination by the presence of older embryos due to maternal egg holding (Harrison et al. 2011), I stained the collections with DAPI and manually removed out-of-stage embryos by suctioning with a pipette tip under the microscope (K. Chen et al. 2013) (Figure 2-4A). Originally, Chen et al (2013) reported 30 dual promoters based on Pol II ChIP-seq occupancy
differences during the MBT. However, after re-analysis of this data and filtering active promoters based on CAGE-seq embryo data (modEncode, see methods), we ended up with 16 dual promoters.

Consistent with the published data, Pol II ChIP-nexus data revealed a clear and progressive accumulation at the pausing position as embryos go from pre-MBT to post-MBT (Figure 2-5B, top). By contrast, this phenomenon is not observed for TBP, whose main footprint occurs upstream of the TSS at all different time points, with no apparent accumulation of signal at the pause site as it happens with Pol II (Figure 2-5B, bottom). In addition, quantification of the downstream vs upstream occupancy ratio showed a progressive increase for Pol II, but not TBP (Figure 2-5C). Overall, this data is consistent with our previous analysis and further suggests no association between TBP and the paused polymerase.

Importantly, I observed overall very little downstream TBP occupancy at the dual promoters compared to what it is seen at TATA or TATA-less promoters in Kc167 cells. To test that the downstream footprint of TBP was not an artifact of the cultured cells, I picked the 1000 promoters with the highest TBP signal in post-MBT embryos and looked at its average binding profile. Consistent the cell culture data, clear TBP footprints were observed both upstream and downstream of TSS (Figure 2-6A). Moreover, to confirm that as in kc167 cells, differences in TBP binding profile correlate with core promoter element sequence composition, I sorted the top 1000 promoters based on upstream vs downstream occupancy ratio, selected the top and bottom 30% and calculated which DNA motifs were enriched at each group. As expected, promoters with a high ratio of downstream/upstream TBP occupancy were depleted in TATA, but enriched in Inr, MTE, DPE and PB, whereas the opposite was true for promoters with a higher proportion of TBP upstream of the TSS (Figure 2-6B).
Taken together, these data confirm the results observed in the kc167 cells and show that while also in the embryo the downstream occupancy of TBP correlates with promoters that experience longer Pol II pausing, these two events appear to occur at different times during the transcription cycle.

Figure 2-6. Downstream TBP does not associated with the paused Pol II
(A) Schematic drawing depicting the collection times used for ChIP-nexus experiments during MBT
(B) The average ChIP-nexus binding profile of TBP and Pol II at dual genes before, during and after MBT shows no correlation between the location of TBP and the paused Pol II. (C) The Pol II, but not TBP, downstream vs upstream occupancy ratio increases progressively from pre-MBT to post-MBT (Wilcoxon test).
Figure 2-7. Downstream TBP is also present in the Drosophila embryo

(A) TBP occupancy at the top 1000 promoter with the highest TBP signal sorted based on the downstream/upstream signal ratio. (B) TBP and Pol II binding profiles at the top and bottom 30% of sorted promoters. (C) Quantification of the downstream/upstream TBP signal ratio at the top and bottom 30% of sorted promoters (Wilcoxon test). (D) Promoter element enrichment analysis that promoters with high TBP upstream enriched for TATA-box while promoters with high TBP downstream enriched for DPE and other downstream elements.
2.6 TFIID is released from the promoter during Pol II pausing

After concluding that TBP does not associate with the core promoter during pausing, we asked whether other TAFs could be present at the promoter at the same time that the paused Pol II. In fact, a functional role enabling promoter-proximal pausing has been proposed for the TAFs (Fant et al. 2020), in particular TAF1 and TAF2, whose DNA-promoter contacts has been proposed to clash with those of the paused polymerase (Louder et al. 2016). However, it is unclear whether the paused Pol II and the TFIID complex can co-exist in the same piece of DNA.

To address this question, I used genome-wide measurements of Pol II pausing half-lives previously generated by our lab (Shao and Zeitlinger 2017) to select promoters with either short ( < 6 min) or stable ( > 27 min) pausing and compared the average binding profile of TFIID (example in Figure 2-7A) as well as the total occupancy (Figure 2-7A and B). In addition, a number of initiation (TFIIA, TFIIB and TFIIF) and pausing (Pol II and Nelf) factors were included as control. This analysis showed that the occupancy of TAFs was significantly higher at promoters with short Pol II pausing half-lives compared to those where pausing is stable. In agreement with this, TFIIA, TFIIB and TFIIF occupancy, were also higher at promoters with short pausing, whereas the occupancy of Pol II and Nelf was higher at stably paused promoters. Altogether, these data indicate that TFIID correlates better with initiation than with pausing and suggests that the holo-complex is released from the promoter DNA when Pol II is stably paused.
Figure 2-8. TFIID occupancy correlates better with initiation than pausing

(A) Comparing the average ChIP-nexus profiles of several TAFs with that of TFIIA and Pol II at shows that TFIID occupancy correlates better with initiation (red) than with pausing (blue). (B) Comparison of the total TFIID occupancy (Wilcoxon test) confirms that like TFIIA, TFIIB and TFIIF, TFIID levels are lower at stably paused promoters (red). In contrast, the levels of pausing factors such as Nelf or Pol II itself correlate with longer Pol II pausing half-lives (blue).
An alternative approach to study the stability of TFIID during the transcription cycle is to measure the changes in binding after triptolide treatment. As explained earlier, triptolide inhibits the helicase activity of TFIIH, halting the formation of a competent transcriptional bubble and impeding promoter escape by Pol II (Titov et al. 2011; Bensaude 2011). If TFIID is quickly released from the promoter once the PIC has been recruited, we would expect an increase in binding after triptolide treatment. By contrast, if TFIID binding becomes somehow stabilized after initiation, we should observe decreased in the total occupancy under the same conditions.

To test this hypothesis, I calculated the relative changes in TFIID binding before and after triptolide treatment at all active promoters. Once again, TFIIA, TFIIB, TFIIF, Pol II and Nelf were used as references. This data revealed an increase in the occupancy for the majority of TFIID subunits as well as for TFIIA, TFIIB and TFIIF, but not for Pol II and Nelf, which experience a severe decrease in binding (Figure 2-8A and B). Notably, the increase in binding after triptolide was quite modest for several subunits (e.g., TAF8, TAF4, TAF3) and no change or even a slight decrease was observed for others (e.g., TAF10, TAF12, TAF13). In addition, the decrease seen for TBP is due the loss of its downstream binding after TRI treatment. Overall, the data suggests that some members of TFIID are more stably associated with chromatin than others across the transcription cycle. However, because triptolite is not able to distinguish between distinct post-initiation events (pausing vs re-initiation) and based on the previous analysis at stably paused promoters, it is unlikely that such more stable association occurs during pausing, but rather within bursts of transcription. This suggest that some subunits/modules of TFIID may remain associated to promoter DNA after Pol II recruitment to allow quick re-initiation. It is also interesting to note that TAF11 undergoes a great increase after triptolide that is not seen for TAF13. This is unexpected since as mentioned before these subunits form a stable heterodimer. Unfortunately, I do not have a reasonable explanation for this phenomenon.
Quite interestingly, comparing the changes after triptolide at TATA vs TATA-less promoters showed a much higher increase of TFIIA, TFIIB and TFIIIF and TATA-less promoters with respect to TATA promoters, but the opposite was seen for the TAFs, which showed a higher increase at TATA promoters vs DPG. While these changes were not statistically significant, the trend seen in the data is consistent with our proposed model in which the PIC being more stable at TATA promoters and TFIID at promoters that lack TATA but contain other elements like Inr, MTE, DPE or PB.
Figure 2-9. TFIID binding is stabilized during PIC assembly

(A-B) The relative occupancy of TFIID after triptolide treatment shows that most subunits are stabilized during PIC formation, although this increase is modest compared to other PIC components such as TFIIB and TFIIF. By contrast, factors known to be more stable during pausing such as Pol II and Nelf showed a sharp decrease in occupancy after triptolide. Notably, the decrease seen for TBP is due to the loss of its downstream peak after triptolide and not due to a role in pausing.
Chapter 3 Discussion

3.1 Genome-wide composition of the TFIID complex

Many studies performed from yeast to mammals have shown that TFIID is recruited to most Pol II promoters and is essential for their transcription. However, how TFIID binds to distinct promoters or whether partial TFIID assemblies regulate transcription of specific gene classes is unclear. For example, early studies in yeast showed that inactivation of individual TAFs did not affect overall mRNA levels, suggesting that each TAF was only required for the expression of a subset of genes. By contrast, recent studies showed that nascent RNA levels are globally reduced after TAFs or TBP depletion suggesting a broad and non-redundant role in gene expression for the different TFIID subunits (Warfield et al. 2017; Fant et al. 2020). On the other hand, various studies in higher eukaryotes have shown that TBP and TAFs are absolutely essential for embryonic development, but not survival of differentiated cells where knock-out or conditional deletion of individual TFIID subunits affected a small percentage of genes. How and why these transcription regulatory differences occur is not completely understood.

The data presented in this thesis work showed that all TFIID subunits appear to be present at active promoters, supporting a general requirement of the complex for Pol II transcription. While I observed differences in the occupancy of individual TAFs at distinct promoter types (e.g., Regulated vs Housekeeping or TATA vs Inr-DPE/PB), in light of the recent functional studies suggesting a universal dependency on TFIID for gene transcription, it is reasonable to interpret that such differences are likely the reflect of distinct rate-limiting steps in TFIID recruitment, which will might be affected depending on the specific features of each promoter type (Bhuiyan and Timmers 2019). These could include 1) the number of enhancers/activators that target the promoter; 2) the presence of distinct core promoter elements (e.g., TATA-box, Inr, DPE...); 3) the chromatin state: levels of active/repressive
histone marks, nucleosome occupancy and positioning; and 4) the presence of negative transcription regulators like NC2 and Mot1, which modulate the DNA-binding dynamics of TBP. The implications of these features in TFIID recruitment will be discussed in more detail below.

A requirement of the holo-TFIID complex for transcription is also intuitive from the structural and functional knowledge of TFIID, where the distinct subunits play specific roles within the complex: maintaining integrity, preventing TBP spurious interactions and loading on DNA, recognition of DNA sequences and interaction with activators and histones modifications. However, cases where a priori essential TFIID subunits are missing, but Pol II transcription is preserved have been reported. For example, in murine cells inactivation of TBP by homologous recombination arrested Pol I and Pol III transcription, but Pol II remained in a transcriptionally active phosphorylation state. Moreover, it was recently shown that a homozygous TAF8 mutation in a patient with intellectual disability resulted in undetectable TAF8 protein and TFIID disassembly, but Pol II transcription seem largely unaffected. Overall these data indicate that while TFIID is essential for Pol II transcription, not having the holo-complex may be less harmful than previously thought. This may be achieved by 1) redundancy of some subunits within the complex, which could allow for partial TFIID assemblies that are still enough to support transcription; 2) the presence of TBP and TAF paralogs that can incorporate within TFIID in the absence of canonical subunits; or 3) residual but undetected levels of the mutated/inactivated proteins were still present in those studies and were sufficient for holo-TFIID assembly.

On the other hand, the fact that we observe TFIID binding occupying all promoters does not imply that all subunits are required for transcription of all genes. However, if there is a distinct genetic requirement of TAFs for the expression of different genes, this cannot be explained by the specific recruitment to distinct promoter types since the entire TFIID complex is recruited to all promoters. We recently proposed the handyman principle to
explain this discrepancy for the coactivator SAGA. In this sense, while a handyman has many tools, he will hardly use them all at the same time, and which tool he uses will depend on what requires maintenance.

3.2 Genome-wide spatial organization of the TFIID complex

To investigate the genome-wide organization of TFIID we mapped the position of each subunit at the top 1000 with the highest signal of TAF1. The position of the ChIP-nexus footprints matched remarkably well the spatial organization of the TFIID complex seen in the recent promoter-bound yeast and human cryo-EM structures, and emphasize the capacity of genome-wide high-resolution techniques to investigate the DNA-bound organization of transcriptional complexes. In this sense, one of the most curious aspects observed in the data is that several subunits from the upstream module (TAF4, TAF10, TA12 and partially TAF5) showed broad, but robust footprints upstream of the TSS that extended to the promoter-proximal region. This promoter region usually works as a binding platform for activators and coactivators, especially (but not exclusively) at housekeeping genes, whose regulatory sequences lie within and immediately upstream of the core promoter (Zabidi et al. 2015). Since it is structurally unlikely that TFIID is able to reach the proximal promoter region by itself, it is reasonable to speculate that broad footprints are due to protein-protein interaction between these TAFs and TFs that are sitting at this region. Importantly, while contacts between the basal transcription machinery and gene-specific activators are believed to be mostly entailed by the Mediator complex, various *in vitro* and *in vivo* studies from yeast to human indicate that TFIID is also able to directly interact with transcriptional activators (Hoey et al. 1993; Ferreri et al. 1994; Kashanchi et al. 1994; Uesugi and Verdine 1999; Mencía et al. 2002; Munz et al. 2003; Hamard et al. 2005; Liu et al. 2009; W.-Y. Chen et al. 2013; Yang and Chang 2013).

Another intriguing aspect was that the main footprint of TBP was located at the PIC
site and not at its cognate motif – the TATA-box. Because this footprint is identical and occurs at the same position of TFIIB and TFIIF (Shao and Zeitlinger 2017), it is reasonable to believe that it comes from interaction with those factors. This is interesting as it could mean that after the initial TFIIA-TBP-TATA interactions, the PIC and TBP itself may no longer rely as much on the TATA-box for stable binding, suggesting that minor rearrangements may occur during PIC formation. Interestingly, TAF13, which forms a heterodimer with TAF11 and can directly interact with TBP (Gupta et al. 2017), showed a clear footprint right where the TATA-box is located, although careful analysis of the TAF13 binding profile showed that such binding peak was preferentially found at promoters that lack a TATA-box (data not shown).

Several genetic and biochemical studies have suggested that the TAF11/13 heterodimer can stabilize the ternary TFIIA-TBP-DNA complex and imparts subtle conformation changes to TFIIA and TBP at the promoter (Lavigne et al. 1999; Kraemer et al. 2001; Robinson et al. 2005). What’s more, using electrophoretic mobility shift assay (EMSA) one of these studies showed that TAF11/13 enhances the stability of TFIIA-TBP-DNA especially at suboptimal promoters harboring a non-canonical TATA-boxes (Robinson et al. 2005). However, more recent structural and biochemical approaches suggest that the TAF11/13 heterodimer compete with TFIIA and DNA for TBP binding and that TAF11/13 are released from interactions with TBP as soon as this one engages promoter DNA (Gupta et al. 2017). Presumably, this releases interaction surfaces within TBP necessary for the recruitment of TFIIA and TFIIIB (Patel et al. 2018), although the fate of TAF11/13 within TFIIID after this occurs is unknown. Because I observe clear ChIP-nexus footprints for TAF11/13 even when PIC assembly is blocked using triptolide, it is likely that these subunits remain engaged at the promoter after they are released from TBP, probably through contacts with other TAFs, where they could play a role modulating the DNA-binding of TBP and
TFIIA, as suggested by the early *in vitro* studies.

Another aspect worth mentioning about TAF11/13 is that while they form a stable heterodimer that is not likely to be spatially separated within TFIID, they are part of different binding modules based on our ChIP-nexus data. The simplest explanation for this is that within the upstream module, TAF11, but not TAF13, makes direct contact with TAF6 and TAF8, which based on the structural data interact with downstream DNA through TAF2 ([Louder et al. 2016; Patel et al. 2018; Kolesnikova et al. 2018](#)). Consistent with this idea, the recent cryo-EM structure of yeast TFIID showed that the TAF11/13 heterodimer could be well-fit (although not unambiguously) within the upstream promoter-bound lobe B and close to the HFD of TAF6-9 ([Kolesnikova et al. 2018](#)). However, the same space was occupied by the HFDs of the TAF8-TAF10 pair in human TFIID ([Louder et al. 2016](#)). Further structural and functional approaches will be necessary to dissect the role and fate of TAF11/13 during transcription initiation and in different species.

Our ChIP-nexus data also confirms that TAF1/7 and TAF2 provide the main contacts with the downstream promoter region encompassing elements like MTE, DPE and PB, as opposed to TAF6/9, which for a long time were the textbook factors to recognize these elements ([Vo Ngoc et al. 2019](#)). The MTE and DPE function cooperatively with the Inr element and can also synergize each other ([Lim et al. 2004; Theisen et al. 2010](#)). In fact, a study of the *Drosophila* downstream core promoter region suggested that the MTE and DPE could be considered as a single unit that provides multiple contacts for TFIID binding ([Theisen et al. 2010](#)). In the ChIP-nexus data, I observed that TFIID makes strong contacts at three well defined positions from near the TSS to the DPE regardless of the promoter type, supporting a unified view of the downstream core promoter region, where key subregions working together seem to be required for efficient TFIID binding.

Besides, we found TAF3 also to be located in the downstream module and its binding profile to be almost identical to the TAF1-2-7 subunits. While TAF3 is part of the mobile
lobe A in the human TFIID structure, its location within the complex after promoter recognition is unclear. In this sense, our data suggests that after relocation of lobe A, TAF3 remains close to the downstream promoter region by interaction with TAF1/TAF7 and/or TAF2, a model consistent with recent crosslinking mass spectrometry (CXMS) analysis in yeast TFIID, which showed abundant unique inter-protein crosslinks between TAF3 and TAF1/7 and TAF2 (Kolesnikova et al. 2018). Besides, TAF3 interacts with the H3K4me3 histone mark found in the +1 nucleosome of active promoters (Vermeulen et al. 2007). Therefore, it is reasonable to think that TAF3 lingers around the downstream core promoter region in order to maintain contacts with the +1 nucleosome. However, I do not observe a direct correlation between TAF3 and H3K4me3 or nucleosome occupancy, suggesting that either these interactions are very transient or that they are no longer required once TFIID is efficiently engaged to the promoter.
3.3 TFIID binding during the transcription cycle and consequences for the transcription dynamics at different promoter types

Earlier, I discussed that while our ChIP-nexus data suggests that all TFIID subunits are present at all active promoters, promoter specific differences in the occupancy of TBP and several TAFs could be interpreted as different mechanisms by which promoters regulate their transcription kinetics through TFIID recruitment. For example, the low occupancy seen for TBP and most TAFs as well as Pol II and TFIIB at the dispersed/housekeeping genes suggest that the whole PIC is overall less stable at these promoters. One could argue that this low binding is just an effect of the spread signal due to the dispersed nature of initiation at housekeeping promoters, however, while the visualization of binding profiles gets heavily impaired by dispersity, the distribution of the total binding signal should not be affected if (as we did) the binding occupancy is calculated in a large-enough window that accounts for this phenomenon.

The simplest explanation to lower occupancy of TFIID at housekeeping promoters with respect to focused/regulated promoters is the lack of bonafide TFIID recognition elements like the TATA-box, Inr or MTE/DPE/PB. Although housekeeping promoters are also enriched for DNA-specific sequences such as DRE, Motif1, Motif6 and Motif7 (Zabidi et al. 2015; Haberle and Stark 2018; Haberle et al. 2019), these are not likely to directly interact with TFIID. This is at least the case for the DRE and Motif1 sequences in Drosophila, which are respectively recognized by the DNA replication-related element (DREF) (Hirose et al. 1993) and the Motif1 binding protein (M1BP) (Li and Gilmour 2013; Baumann and Gilmour 2017), indicating that specific TFs other than TFIID recognize these elements in vivo.

The observation that TFIID and the PIC are unstable at housekeeping promoters does not imply low transcriptional levels or low TFIID recruitment. In fact, housekeeping genes show robust levels of expression in our data and TFs like DREF and M1BP are likely to
increase TFIID recruitment either by direct interaction with the TAFs or indirectly through contacts with the Mediator complex. Housekeeping promoters also present strong positioning of the +1 nucleosome (Rach et al. 2011) and are marked by H3K4me3 and H3K27ac (Haberle and Stark 2018), which can be recognized by the PHD finger of TAF3 (Vermeulen et al. 2007) and the double-bromodomain of TAF1 (Jacobson et al. 2000; Huisinga and Pugh 2004), respectively, and might help increase the rate of TFIID recruitment to these promoters. Moreover, a novel histone modification, serotonylation of glutamine-5 in histone H3 (H3Q5ser), has been shown to enhance TFIID binding at human promoters, although the molecular mechanism and the TAFs involved remain to be determined (Farrelly et al. 2019).

High recruitment of TFIID at promoters but low DNA stability is consistent with the non-bursty, constitutive mode of transcription (Poisson) attributed to housekeeping genes (Zenklusen et al. 2008; Larsson et al. 2019). A tempting model could be that at these promoters, the PIC may be able to recruit just one or two polymerases before it disassembles due to the lack of bonafide elements that position and stabilize TFIID binding (Figure 4-1). However, robust recruitment of the TFIID complex could be achieved by the presence of promoter-proximal activator-specific sequences as well as the maintenance of high levels of the H3K4me3 and H3K27ac at the +1 nucleosome. Besides, this quick recruitment of TFIID by proximal TFs and histone marks may prevent avid nucleosomes from occupying the core promoter region, thus helping maintain a continuously active transcriptional state (Workman and Roeder 1987; Meisterernst et al. 1990).

In contrast to housekeeping promoters, we observed high occupancy and sharp footprints for TFIID and the PIC (TFIIA,TFIIB, TFIIF) at focused/regulated promoters. However, dividing these promoters into discrete groups based on motif composition revealed a clear duality in TFIID and PIC binding between TATA and TATA-less promoters. First, we found the occupancy of upstream TBP, TFIIA, TFIIB and TFIIF, but not TAFs to be
higher at TATA promoters. We interpret these results as the PIC being more stable at promoters in which a TATA-box is present, consistent with recent single-molecule footprinting experiments in which the region for PIC assembly showed strong protection against DNA-methylation at TATA promoters, but not at TATA-less promoters (Krebs et al. 2017).

Assuming that the PIC is more stable at TATA promoters, the fact that TAFs occupancy was lower there than at TATA-less promoters suggests that these subunits are not required to remain associated with the DNA to sustain initiation once the PIC is established. We speculate that after promoter recognition and PIC assembly, TAFs are released from DNA by the transcribing Pol II, which as proposed by recent structural data, clashes with TFIID at the downstream promoter region (Patel et al. 2018). Our data support this hypothesis as comparing TFIID occupancy at promoters with either very short or very high Pol II pausing half-life revealed that binding of all subunits was reduced when Pol II is stably paused. Besides, these results are consistent with the overall increased binding for most TAFs after triptolide treatment. Triptolide inhibits promoter escape by Pol II by halting the formation of a competent transcriptional bubble (Titov et al. 2011; Bensaude 2011). Presumably, because Pol II cannot move forward into the downstream promoter, clashes with the TAFs located in this region would be prevented, resulting in TFIID stabilization. In agreement with our triptolide data, an increased in TAF1 and TAF4 occupancy was observed in yeast after depletion of the TFIIDH subunit kin28, which also inhibits Pol II promoter escape, but by preventing Pol II-CTD Ser5 phosphorylation, required for the recruitment of factors involved in elongation and other associated processes (Knoll et al. 2020).

Although several lines of evidence, including ours, suggest that TAFs are released from promoter DNA as Pol II begins transcribing, other models have been proposed. In particular, recent work by the Buratowski lab suggested that in yeast, TAF1/7 remain stably associated with the downstream promoter region to allow activator-independent re-initiation
(Joo et al. 2017). To favor this model, one could interpret that the lower TAFs ChIP-nexus occupancy seen in steady-state conditions vs after triptolide treatment or after kin28 depletion just reflects weaker DNA-contacts when Pol II passes through the downstream promoter region, but does not necessarily mean that TAFs are fully dissociated from DNA. Maybe, TAFs are able to remain bound to the non-template DNA while Pol II is transcriptionally engaged at the downstream promoter region, resulting in lower ChIP efficiency and therefore in less measured occupancy. While a model like this may be consistent with early studies in vitro that suggested that promoter-bound TBP-TFIID complexes were long-lived (30-45 min) (Workman and Roeder 1987; Timmers and Sharp 1991; Hoopes et al. 1998), it is in sharp contrast with more recent live-imaging approaches in vivo, which have shown that the promoter binding of TFIID and other BTFs is highly dynamic, with shorter DNA residency-times ranging from seconds to few minutes (Sprouse et al. 2008; de Graaf et al. 2010; Revyakin et al. 2012; Zhang et al. 2016; Presman et al. 2017; Wu et al. 2020), a discrepancy that can be explained by the presence of transcriptional factors that regulate PIC binding dynamics and which are absent in vitro. Therefore, it is unlikely that TAFs remain stably bound to promoter DNA during the transcription cycle.

Another recent model was proposed by Dylan Taatjes based on in vitro and in vivo experiments in fly (Fant et al. 2020). Here, TAFs would be transiently released from the downstream promoter as Pol II goes through, but the TAFs in the upstream module could still remain associated with the promoter, allowing quick re-binding of the downstream TAFs once Pol II makes it into productive elongation. While this is a tempting model, there are several things to consider. First, this work does not provide direct or indirect evidence about the binding of upstream TFIID subunits. Second, a model in which upstream TAFs remain bound to the DNA to allow quick reloading of downstream TAFs and enhance re-initiation also implies stable TFIID DNA binding, which as discussed above is inconsistent with the low residency times seen in live-imaging experiments. Third, our ChIP-nexus data
Discussion

showed that occupancy of all TFIID subunits, including the upstream TAFs, was lower at stably paused promoters, arguing against an association between TFIID and pausing. Interestingly, after blocking Pol II promoter escape with triptolide, we observed that the occupancy of most TAFs located in the upstream module was basically unchanged. This suggests that the binding of the TAFs within the upstream module may in fact be more stable than those in the middle or downstream module, to allow quick re-binding. However, if this is true, it is more likely to occur within bursts of transcription and not when the Pol II is paused, which can last for minutes (Core and Adelman 2019) and it is thought to occur between bursts, when the promoter is not active (Shao and Zeitlinger 2017; Pimmett et al. 2021). Thus, the modular nature of TFIID may allow quick re-initiation within bursts of transcription.

It is important to highlight that the fact that TFIID is not bound to the promoter during Pol II pausing does not mean that it cannot be functionally involved in the process, as recently proposed (Fant et al. 2020). As it happens, all the promoter elements that are recognized by the TAFs in the downstream module (Inr, MTE, DPE/PB) correlate with Pol II pausing (Hendrix et al. 2008; Shao and Zeitlinger 2017; Shao et al. 2019). The same way these elements may help quick and precise re-binding of TAFs, they could also facilitate pausing in higher organisms, as the anchoring of TAFs to downstream core promoter region may work as a physical barrier that slows down the early transcribing Pol II. Consistent with this idea, fast depletion of TAF1 in Drosophila S2 cells resulted in increased accumulation of transcribing polymerases downstream of the TSS that failed to enter productive elongation (Fant et al. 2020). One possible scenario that reconciles the existent evidence is that in steady-state conditions and within a burst of transcription, pausing is rapidly overcome by the action of activators and positive elongation factors like PTEF-b. However, if any of the rate-limiting factors required to “push” Pol II out of the promoter into productive elongation fails, Pol II will be enter into a longer pausing state initially stablished by the combined
action of promoter sequences and TFIID, but further stabilized by the action of the bonafide pausing factors DSIF and NELF. This model is in agreement with a very recent live-imaging study in Drosophila embryos showing that Pol II enters stable pausing in a stochastic rather than obligatory manner during transcription (Pimmett et al. 2021).

Another important aspect that can be extracted from our data is that despite the differences in TAFs occupancy between TATA and TATA-less promoters, their binding organization remained consistent. Earlier reports have proposed that the DNA-bound conformation of TFIID may differ depending on the core promoter sequence. However, those models were very speculative as they were drawn based on comparisons between in vitro photo-crosslinking experiments of TAF1 (Lee et al. 2005) and TAF6/9 (Burke and Kadonaga 1997) at a small set of promoters in human and fly, respectively, and on differences among whole-chromosome TAFs binding profiles from low-resolution ChIP-chip data in yeast (Ohtsuki et al. 2010). Thus, there is really no direct evidence that supports the existence of alternative TFIID configurations depending on core promoter sequence. Rather, our data suggest that core promoter sequences regulate their transcription dynamics by affecting the kinetics of TFIID recruitment and not its DNA-bound conformation.

In contrast to the TAFs, TATA and TATA-less promoters showed drastic differences on its TBP binding profile. From the beginning, it was tempting to speculate that the downstream TBP footprint was related to the canonical state of TFIID seen in structural studies when the complex is in solution (Cianfrocco et al. 2013; Louder et al. 2016; Patel et al. 2018). However, the downstream core promoter region is a place where the binding of different factors converge at different stages during the transcription cycle and it is therefore possible that this footprint may be arising from interactions other than those occurring before the TBP relocation. Therefore, there may be alternative explanations for the downstream footprint of TBP and, thus, I first set out to test alternative models.

First, I reasoned that the downstream binding of TBP could arise from the TAFs
contacts with the downstream promoter region during PIC assembly. However, our data revealed that downstream TBP binding was abolished almost completely after the addition of triptolide, and that a shift occurs between the upstream and downstream occupancy of TBP compared to the control conditions. The fact that downstream TBP and TAF1 occupancy showed opposite behaviors when the PIC was stabilized by inhibiting Pol II promoter escape indicates that TBP is not located downstream of the TSS during PIC formation, because, if such footprint reflected contacts through the downstream TAFs, then TBP occupancy should also have increased at this location, but that was not the case.

Second, I also had the suspicion that the downstream footprint of TBP could be related to the paused Pol II. There were several reasons to believe this: 1) The downstream TBP and paused Pol II footprints occurs roughly at the same site; 2) downstream TBP is present at paused promoters, which are enriched in motifs like Inr, MTE, DPE or PB, recognized by TFIID (Shao and Zeitlinger 2017; Shao et al. 2019); and 3) a direct role in Pol II pausing has been recently attributed to TFIID (Fant et al. 2020). However, the TBP and Pol II triptolide data strongly argued against downstream TBP occurring during Pol II pausing. These data therefore suggest that TBP is found at the downstream location neither during PIC formation nor during Pol II pausing.

I reasoned that the most likely explanation is that this binding footprint belongs to a pre-loading state of TBP, as proposed based on the cryo-EM human TFIID data (Patel et al. 2018). Notably, such pre-loading, TBP-inhibited conformation is not observed once TFIID binds promoter DNA in the structural studies, as TBP quickly relocates to upstream promoter region to scan for the TATA-box. However, based on our data, it seems that either this state is more stable than previously thought or it occurs frequently enough that it can be captured on DNA at both TATA and TATA-less promoters.

Importantly, we cannot exclude from this data that the footprints of TBP at the downstream promoter are due to direct DNA binding. In fact, TBP binds DNA with low
preference between TATA or TATA-less sequences. However, DNA binding by TBP is highly regulated *in vivo* by different factors (TAF1, TAF11/13, NC2/BTAF1), which ensure no spurious TBP binding at random sites. Moreover, the downstream core promoter region is most often bound by TFIID or the paused Pol II and is therefore unlikely that TBP can bind stably enough to this position to result in such strong footprint.

The observation that at TATA-less promoters, the occupancy of early PIC factors like TFIIA, TFIIB and TFIIF is low while for TAFs is high, also supports a model for TBP pre-loading state. Since this state presumably precedes formation of the PIC, only TAFs should be present at the core promoter when TBP is being pre-loaded downstream of the TSS. Certainly, because Pol II pausing is more stable at TATA-less promoters, one could argue that the low levels of PIC could just be due to fewer initiation events. However, TFIIA, TFIIB and TFIIF undergo a robust increase in occupancy after triptolide treatment also at TATA-less promoters, suggesting that the low binding observed in control conditions is due to an unstable PIC rather than just scarce initiation. Besides, our data also goes in line with the recent single-molecule tracking data in yeast, in which rapid depletion of TBP from the nuclei using the anchor-away system (AA) abolished chromatin association of all PIC components except for TAFs and mediator, which experienced a great increase in their residency time (Wu et al. 2020).

The differences in TFIID binding seen between TATA-containing and TATA-less promoters may provide mechanistic insights into how these promoters control their transcription kinetics. At both types of promoters, TBP seems to be pre-loaded downstream of the TSS by TFIID and then relocated upstream, where recruitment of the remaining PIC components will occur. However, at TATA-promoters, a more stable PIC may facilitate the loading of a large number of consecutive polymerases resulting in a strong burst of transcription. Such transcriptional burst may last until the PIC disassembles from the core promoter, potentially due to the action of negative regulatory factors like NC2 and BTAF1.
(van Werven et al. 2008). This interpretation is consistent with several studies in different organisms showing a clear role for the TATA-box increasing burst size, presumably reflecting the number of polymerases loaded in one single activation event. On the other hand, the transcriptional activity of TATA-less promoters seems to be dominated by the burst frequency (number of bursts that occur in particular period of time), while burst sizes tend to be much smaller than for TATA promoters (Larsson et al. 2019). Notably, this should not be interpreted as TATA-less promoters does not burst at all, but rather as less polymerases could be loaded from the same PIC due to more unstable binding.

These potential mechanistic differences become even more interesting when considering the functional role of TATA and TATA-less promoters. TATA-promoters are mostly active in terminally differentiated adult cells and their expression does not seem to require coordinated transcription across the cells within the same tissue, consistent with the high expression noise seen for these promoters in smRNAfish, live-imaging and single-cell RNA-seq experiments from yeast to human (Zoller et al. 2015; Blake et al. 2003; Tiross and Barkai 2008; Lehner 2008; Carey et al. 2013; Raser and O’Shea 2004; Murphy et al. 2010; Blake et al. 2006; Faure et al. 2017; Ramalingam et al. 2020). Conversely, core promoters depleted in TATA but enriched in elements like Inr, DPE and PB often drive the expression of genes involved in early developmental processes such as patterning and morphogenesis (Engström et al. 2007) and therefore need to be synchronously activated during embryogenesis as well as to precisely respond to different gradients of concentration from gene-specific activators. This model fits with recent quantitative imaging of transcription dynamics using live Drosophila embryos, showing that short periods of inactivity (~20 seconds) occurred between each initiation event at a developmental TATA-less paused promoter, but not at TATA or TATA-like containing promoters (Pimmett et al. 2021). Therefore, it is tempting to speculate that the absence of TATA-box may allow tighter regulation of transcription by fine control of TFIID loading onto the promoter, as opposed
to a stable PIC where a greater and more variable number of polymerases will be loaded each burst. In fact, several studies have reported a decrease in transcriptional noise upon TATA-box mutations (Raser and O’Shea 2004; Blake et al. 2006; Murphy et al. 2010; Hornung et al. 2012).

Within this context, it is interesting to note that yeast promoters, which primarily respond to physiological cues, are enriched in A-T sequences, whereas TATA-less promoters predominate in metazoans like flies or vertebrates, where most genes need to respond to developmental signals (Tora and Timmers 2010; Juven-Gershon and Kadonaga 2010; Rhee and Pugh 2012b). Thus, it possible that the core promoter regions of higher organisms, especially those that require precise and coordinated expression, may have evolved to regulate transcription by the frequency of activation, which presumably involves de novo TFIID loading, allowing finer control of mRNA production within the cell. A model for the transcription dynamics of three main types of promoters is shown in Figure 4.
Figure 3-1. Model for TFIID binding dynamics at different promoter types
Chapter 4 Conclusions & future directions

Previous *in vivo* studies trying to understand promoter recognition by TFIID often employed low resolution techniques like ChIP-seq and focused on one or few subunits of the complex, with little effort dedicated in dissecting changes within TFIID architecture or composition between promoters types and its relationship with other PIC components. In this thesis, I set out to fill this gap by analyzing the binding at high resolution of the fourteen TFIID subunits and other basal transcription factors at distinct core promoters *in vivo*.

From this work, it seems that all subunits TFIID bind to active promoters and that differences in occupancy between promoters types likely reflect distinct rate limiting steps in TFIID recruitment rather than partial assemblies of the complex. Moreover, we found that the *in vivo* ChIP-nexus footprints divide TFIID into three sub-modules whose organization is remarkably homogeneous across different promoter types. In contrast, the binding profile of TBP is promoter-specific and served as the basis for identifying TATA, pausing and housekeeping promoters and their underlying sequence elements. Our results also showed that TBP is found upstream and downstream of the start site, but that the amount and exact profile at each position is dependent on the promoter type and transcriptional state. These results are consistent with a model in which TBP is first loaded downstream of the core promoter and then transferred upstream to assemble the transcription machinery. Finally, we found that while TAFs have been implicated in the establishment of Pol II pausing, these are released from promoter DNA once Pol II reaches the pause site.

Although this thesis work shed light on the binding of the holo-TFIID complex, much more work is required to understand the molecular mechanism by which TFIID is recruited to promoters and regulates transcription. First, it would be extremely informative to combine ChIP-nexus with neural network approaches to predict which DNA regulatory elements are the most influential for TFIID binding across the genome. In addition, large scale promoter mutagenesis followed by reporter-ChIP-nexus assays would be extremely useful to validate
Conclusions & future directions

such findings and will help tremendously in uncovering the key DNA features that are involved in TFIID recruitment to promoters.

Second, to demonstrate that the first step of transcription initiation involves pre-loading of TBP downstream of the TSS, it would be crucial to dissect the structural organization of the TFIID complex at distinct endogenous promoters. Based on our model, the pre-loading state may be less transient than previously thought, but this hypothesis is built upon ChIP-nexus data and triptolide treatments. Thus, more functional studies need to be included to understand how this step is regulated. One possible experiment may involve mutations on the TAF1 TAND domain. Since this domain keeps TBP in an inhibited state during the pre-loading state, mutations that increase or decrease its affinity for TBP should result in higher or lower binding of TBP downstream footprint.

Third, live-imaging approaches need to be incorporated to dissect the relationship between TFIID and transcriptional bursts. Several studies have focused on how distinct promoter sequences affect transcription dynamics, but little focus has been put on TFIID, which is the main factor to recognize these sequences. Thus, it would essential to examine transcriptional bursting after inactivating distinct components of TFIID and relate that to changes in promoter sequences to better understand how transcription dynamics are regulated at distinct promoter types. Additionally, these assays would be helpful to elucidate whether specific modules/subunits of TFIID are particularly important for transcription re-initiation within transcriptional bursting. Although our data suggested a release of the whole complex during Pol II pausing, this is thought to occur between and not within burst of transcription. If combined with biochemical assays like TFIID pull-downs from nuclear vs
chromatin fractions, these experiments would help tremendously to shed light on how the modular structure of TFIID relates to transcription dynamics.

I believe that over the next years, these studies will reveal TFIID as a key player in dictating transcription dynamics at distinct promoter types, beyond acting as a mere scaffold that sits at the core promoter to allow PIC assembly.
Chapter 5 Materials & Methods

5.1 Materials

Table 5-1. Cell lines and fly stocks

<table>
<thead>
<tr>
<th>Name</th>
<th>Species</th>
<th>Type</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kc167</td>
<td>Drosophila melanogaster</td>
<td>Cell line</td>
<td>DGRC, stock number #1</td>
</tr>
<tr>
<td>OreR</td>
<td>Drosophila melanogaster</td>
<td>Fly stock</td>
<td>Maintained in house</td>
</tr>
</tbody>
</table>

Table 5-2. Antibodies

<table>
<thead>
<tr>
<th>Target</th>
<th>Host</th>
<th>Type</th>
<th>Antigen</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drosophila TFIIA</td>
<td>rabbit</td>
<td>polyclonal</td>
<td>Full-length</td>
<td>James Kadonaga, UCSD</td>
</tr>
<tr>
<td>Drosophila TFIIB</td>
<td>rabbit</td>
<td>polyclonal</td>
<td>Full-length</td>
<td>James Kadonaga, UCSD</td>
</tr>
<tr>
<td>Drosophila TFIIF</td>
<td>rabbit</td>
<td>polyclonal</td>
<td>Full-length</td>
<td>James Kadonaga, UCSD</td>
</tr>
<tr>
<td>Drosophila RPB3</td>
<td>rabbit</td>
<td>polyclonal</td>
<td>Full-length</td>
<td>GeneScript, #163185-7</td>
</tr>
<tr>
<td>Drosophila NELF-E</td>
<td>rabbit</td>
<td>polyclonal</td>
<td>Full-length</td>
<td>GeneScript, #163185-7</td>
</tr>
<tr>
<td>Drosophila TBP</td>
<td>rabbit</td>
<td>polyclonal</td>
<td>Full-length</td>
<td>James Kadonaga, UCSD</td>
</tr>
<tr>
<td>Drosophila TAF1</td>
<td>rabbit</td>
<td>polyclonal</td>
<td>DEGSIGNGLDLTGILC</td>
<td>GeneScript, #U7893CL140-11</td>
</tr>
<tr>
<td>Drosophila TAF2</td>
<td>rabbit</td>
<td>polyclonal</td>
<td>aa 640-971</td>
<td>GeneScript, #U4123BK150-9</td>
</tr>
<tr>
<td>Drosophila TAF3</td>
<td>rabbit</td>
<td>polyclonal</td>
<td>aa 133-319</td>
<td>GeneScript, #U7724CL140-9</td>
</tr>
<tr>
<td>Drosophila TAF4</td>
<td>rabbit</td>
<td>polyclonal</td>
<td>aa 209-383</td>
<td>GeneScript, #U7893CL140-3</td>
</tr>
<tr>
<td>Drosophila TAF5</td>
<td>rabbit</td>
<td>polyclonal</td>
<td>aa 24-135</td>
<td>GeneScript, #U7893CL140-5</td>
</tr>
<tr>
<td>Drosophila TAF6</td>
<td>rabbit</td>
<td>polyclonal</td>
<td>aa 24-135</td>
<td>GeneScript, #U1266BL120-3</td>
</tr>
<tr>
<td>Drosophila TAF7</td>
<td>rabbit</td>
<td>polyclonal</td>
<td>aa 1-181</td>
<td>GeneScript, #U1266BL120-6</td>
</tr>
<tr>
<td>Drosophila TAF8</td>
<td>rabbit</td>
<td>polyclonal</td>
<td>aa 223-328</td>
<td>GeneScript, #U7893CL140-7</td>
</tr>
<tr>
<td>Drosophila TAF9</td>
<td>rabbit</td>
<td>polyclonal</td>
<td>aa 1-188</td>
<td>GeneScript, #529220-34</td>
</tr>
<tr>
<td>Drosophila TAF10</td>
<td>rabbit</td>
<td>polyclonal</td>
<td>aa 1-93</td>
<td>GeneScript, #U7970DB220-5</td>
</tr>
<tr>
<td>Drosophila TAF11</td>
<td>rabbit</td>
<td>polyclonal</td>
<td>aa 1-98</td>
<td>GeneScript, #U1266BL120-9</td>
</tr>
<tr>
<td>Drosophila TAF12</td>
<td>rabbit</td>
<td>polyclonal</td>
<td>aa C-1-36</td>
<td>GeneScript, #U7970DB220-1</td>
</tr>
<tr>
<td>Drosophila TAF13</td>
<td>rabbit</td>
<td>polyclonal</td>
<td>Full-length</td>
<td>GeneScript, #U3982CL190-3</td>
</tr>
</tbody>
</table>
### Table 5-3. Standard ChIP-buffer

<table>
<thead>
<tr>
<th>Compound</th>
<th>Concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tris pH 8.0</td>
<td>10 mM</td>
</tr>
<tr>
<td>NaCl</td>
<td>140 mM</td>
</tr>
<tr>
<td>Triton X100</td>
<td>1%</td>
</tr>
<tr>
<td>Sodium deoxycholate</td>
<td>0.1%</td>
</tr>
<tr>
<td>N-lauroylsarcosine</td>
<td>0.5%</td>
</tr>
<tr>
<td>SDS</td>
<td>0.1%</td>
</tr>
</tbody>
</table>

### Table 5-4. ChIP-nexus wash buffer A

<table>
<thead>
<tr>
<th>Compound</th>
<th>Concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>TE</td>
<td>10 mM</td>
</tr>
<tr>
<td>Triton X100</td>
<td>0.1%</td>
</tr>
</tbody>
</table>

### Table 5-5. ChIP-nexus wash buffer B

<table>
<thead>
<tr>
<th>Compound</th>
<th>Concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tris pH 8.0</td>
<td>20 mM</td>
</tr>
<tr>
<td>NaCl</td>
<td>150 mM</td>
</tr>
<tr>
<td>Triton X100</td>
<td>1%</td>
</tr>
<tr>
<td>Sucrose</td>
<td>5.2%</td>
</tr>
<tr>
<td>EDTA</td>
<td>5 mM</td>
</tr>
<tr>
<td>SDS</td>
<td>0.2%</td>
</tr>
</tbody>
</table>

### Table 5-6. ChIP-nexus wash buffer C

<table>
<thead>
<tr>
<th>Compound</th>
<th>Concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tris pH 8.0</td>
<td>5 mM</td>
</tr>
<tr>
<td>NaCl</td>
<td>250 mM</td>
</tr>
<tr>
<td>Triton X100</td>
<td>0.5%</td>
</tr>
<tr>
<td>HEPES</td>
<td>25 mM</td>
</tr>
<tr>
<td>Sodium deoxycholate</td>
<td>0.05%</td>
</tr>
<tr>
<td>EDTA</td>
<td>0.5 mM</td>
</tr>
</tbody>
</table>

### Table 5-7. ChIP-nexus wash buffer D

<table>
<thead>
<tr>
<th>Compound</th>
<th>Concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tris pH 8.0</td>
<td>10 mM</td>
</tr>
<tr>
<td>LiCl</td>
<td>250 mM</td>
</tr>
<tr>
<td>IGEPAL CA-630</td>
<td>0.5%</td>
</tr>
<tr>
<td>Sodium deoxycholate</td>
<td>0.5%</td>
</tr>
<tr>
<td>EDTA</td>
<td>10 mM</td>
</tr>
</tbody>
</table>

### Table 5-8. Tris buffers

<table>
<thead>
<tr>
<th>Compound</th>
<th>Concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tris pH 7.5, 8.0 or 9.5</td>
<td>10 mM</td>
</tr>
</tbody>
</table>

### Table 5-9. RIPA buffer

<table>
<thead>
<tr>
<th>Compound</th>
<th>Concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>LiCl</td>
<td>0.5 M</td>
</tr>
<tr>
<td>NaCl</td>
<td>50 mM</td>
</tr>
<tr>
<td>IGEPAL CA-630</td>
<td>1%</td>
</tr>
<tr>
<td>Sodium deoxycholate</td>
<td>0.7%</td>
</tr>
<tr>
<td>EDTA</td>
<td>1 mM</td>
</tr>
</tbody>
</table>

### Table 5-10. TE buffer

<table>
<thead>
<tr>
<th>Compound</th>
<th>Concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>EDTA</td>
<td>1 mM</td>
</tr>
<tr>
<td>Tris pH 8.0</td>
<td>50 mM</td>
</tr>
</tbody>
</table>

### Table 5-11. Elution buffer

<table>
<thead>
<tr>
<th>Compound</th>
<th>Concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>EDTA</td>
<td>10 mM</td>
</tr>
<tr>
<td>Tris pH 8.0</td>
<td>50 mM</td>
</tr>
<tr>
<td>SDS</td>
<td>1%</td>
</tr>
</tbody>
</table>
### Table 5-12. Chemicals and reagents

<table>
<thead>
<tr>
<th>Name</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agencourt AMPure XP</td>
<td>Beckman Coulter Genomics #A63881</td>
</tr>
<tr>
<td>Chloroform</td>
<td>Sigma-Aldrich #C2432</td>
</tr>
<tr>
<td>CircLigase™ ssDNA Ligase</td>
<td>Illumina (Epicentre) #CL4115K</td>
</tr>
<tr>
<td>Direct-zol RNA MiniPrep kit</td>
<td>Genesee #11-330</td>
</tr>
<tr>
<td>DMSO</td>
<td>Sigma-Aldrich #276855</td>
</tr>
<tr>
<td>Dynabeads Protein A</td>
<td>Life Technologies #100-08D</td>
</tr>
<tr>
<td>Dynabeads Protein G</td>
<td>Life Technologies #100-04D</td>
</tr>
<tr>
<td>FastDigest BamHI</td>
<td>ThermoFisher Scientific #FERFD0054</td>
</tr>
<tr>
<td>Gibson Assembly® Master Mix</td>
<td>New England Biolabs #E2611S</td>
</tr>
<tr>
<td>HyClone SFX-Insect Cell Culture Media</td>
<td>ThermoFisher Scientific #SH3027802PM</td>
</tr>
<tr>
<td>IBI High Speed Plasmid Mini Kit</td>
<td>MIDSCI #IB47101</td>
</tr>
<tr>
<td>Klenow Fragment (3’ to 5’ exo-)</td>
<td>New England Biolabs #M0212S</td>
</tr>
<tr>
<td>Laemmli Sample Buffer (2x)</td>
<td>Bio-Red #161-0747</td>
</tr>
<tr>
<td>Lambda Exonuclease</td>
<td>New England Biolabs #M0262L</td>
</tr>
<tr>
<td>NEBNext dA-tailing module</td>
<td>New England Biolabs #E6053L</td>
</tr>
<tr>
<td>NEBNext end-repair module</td>
<td>New England Biolabs #E6050L</td>
</tr>
<tr>
<td>NEBNext Multiplex Oligos for Illumina</td>
<td>New England Biolabs #E7335S</td>
</tr>
<tr>
<td>NEBNext Quick Ligation Module</td>
<td>New England Biolabs #E6056S</td>
</tr>
<tr>
<td>Phenol : Chloroform : iso-Amyl alcohol</td>
<td>VWR #97064-824</td>
</tr>
<tr>
<td>Protease Inhibitor Cocktail tablets EDTA-free</td>
<td>Roche Diagnostics Corporation #5056489001</td>
</tr>
<tr>
<td>Proteinase K</td>
<td>Life Technologies #25530-049</td>
</tr>
<tr>
<td>QIAquick Gel Extraction Kit</td>
<td>Qiagen #28706</td>
</tr>
<tr>
<td>Qubit™ dsDNA HS Assay Kit</td>
<td>ThermoFisher Scientific # Q32854</td>
</tr>
<tr>
<td>Quick Ligation Kit</td>
<td>New England Biolabs #M2200L</td>
</tr>
<tr>
<td>recJ</td>
<td>New England Biolabs #M0264L</td>
</tr>
<tr>
<td>RNase A</td>
<td>ThermoFisher Scientific # EN0531</td>
</tr>
<tr>
<td>RNase H</td>
<td>New England Biolabs #M0297S</td>
</tr>
<tr>
<td>T4 DNA Polymerase</td>
<td>New England Biolabs # M0203S</td>
</tr>
<tr>
<td>Triptolide</td>
<td>TOCRIS Bioscience #3253</td>
</tr>
<tr>
<td>TRIzol™ Reagent</td>
<td>ThermoFisher Scientific #15596026</td>
</tr>
<tr>
<td>Name</td>
<td>Type</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>---------------</td>
</tr>
<tr>
<td>Nex_adapter_BNSBamHI_CTGA</td>
<td>Adaptor: barcoded</td>
</tr>
<tr>
<td>Nex_adapter_BNSBamHI_TGAC</td>
<td>Adaptor: barcoded</td>
</tr>
<tr>
<td>Nex_adapter_BNSBamHI_GACT</td>
<td>Adaptor: barcoded</td>
</tr>
<tr>
<td>Nex_adapter_BNSBamHI_ACTG</td>
<td>Adaptor: barcoded</td>
</tr>
<tr>
<td>Nex_cut_BamHI</td>
<td>Oligo for digestion</td>
</tr>
<tr>
<td>Nexus_primer_U</td>
<td>Primer: universal</td>
</tr>
<tr>
<td>Nexus_primer_B01</td>
<td>Primer: barcoded</td>
</tr>
<tr>
<td>Nexus_primer_B02</td>
<td>Primer: barcoded</td>
</tr>
<tr>
<td>Nexus_primer_B03</td>
<td>Primer: barcoded</td>
</tr>
<tr>
<td>Nexus_primer_B04</td>
<td>Primer: barcoded</td>
</tr>
<tr>
<td>Nexus_primer_B05</td>
<td>Primer: barcoded</td>
</tr>
<tr>
<td>Nexus_primer_B06</td>
<td>Primer: barcoded</td>
</tr>
<tr>
<td>Nexus_primer_B07</td>
<td>Primer: barcoded</td>
</tr>
<tr>
<td>Nexus_primer_B08</td>
<td>Primer: barcoded</td>
</tr>
<tr>
<td>Nexus_primer_B09</td>
<td>Primer: barcoded</td>
</tr>
<tr>
<td>Nexus_primer_B10</td>
<td>Primer: barcoded</td>
</tr>
<tr>
<td>Nexus_primer_B11</td>
<td>Primer: barcoded</td>
</tr>
<tr>
<td>Nexus_primer_B12</td>
<td>Primer: barcoded</td>
</tr>
<tr>
<td>Nexus_primer_B13</td>
<td>Primer: barcoded</td>
</tr>
<tr>
<td>Nexus_primer_B14</td>
<td>Primer: barcoded</td>
</tr>
<tr>
<td>Nexus_primer_B15</td>
<td>Primer: barcoded</td>
</tr>
<tr>
<td>Nexus_primer_B16</td>
<td>Primer: barcoded</td>
</tr>
<tr>
<td>Nexus_primer_B18</td>
<td>Primer: barcoded</td>
</tr>
<tr>
<td>Nex_primer_B19</td>
<td>Primer: barcoded</td>
</tr>
<tr>
<td>Nex_primer_B20</td>
<td>Primer: barcoded</td>
</tr>
</tbody>
</table>
5.2 Methods

Cell culture

Drosophila Kc167 cells (purchased from DGRC; negative for mycoplasma contamination) were grown at 25 °C in HyClone SFX-Insect Cell Culture Media. Transcription inhibitors were added directly into SFX media. Cells were treated with 10 μM triptolide (TOCRIS Bioscience, 3253; dissolved in DMSO), 500 nM flavopiridol (Santa Cruz Biotechnology, sc-202157; dissolved in DMSO) or 50 μM DRB (Sigma, D1916; in DMSO) at room temperature for 1 h. Equivalent amounts of DMSO treatment (2%) were used as the control.

ChIP-nexus experiments

For each ChIP-nexus experiment, 107 Drosophila melanogaster Kc167 cells were fixed with 1% formaldehyde in culturing media at room temperature for 10 min with rotation. Fixed cells were washed with cold PBS, incubated with Orlando and Paro’s Buffer (0.25% Triton X-100, 10 mM EDTA, 0.5 mM EGTA, 10 mM Tris-HCl pH 8.0, with freshly added Protease Inhibitor) for 10 min at room temperature with rotation, and then centrifuged and re-suspended in ChIP Buffer (10 mM Tris-HCl, pH 8.0; 140 mM NaCl; 0.1% SDS; 0.1% sodium deoxycholate; 0.5% sarkosyl; 1% Triton X-100, with freshly added Protease Inhibitor). Sonication was performed with a Bioruptor Pico for five rounds of 30 s on and 30 s off. Chromatin extracts were then centrifuged at 16000 g for 5 min at 4°C, and supernatants were used for ChIP.

To couple Dynabeads with antibodies, 50 μl Protein A and 50 μl Protein G Dynabeads were used for each ChIP-nexus experiment and washed twice with ChIP Buffer. After removing all the liquid, Dynabeads were resuspended in 400 μl ChIP Buffer. 10 μg
antibodies were added, and tubes were incubated at 4°C for 2 hr. with rotation. After the incubation, antibody-bound beads were washed twice with ChIP Buffer.

For chromatin immunoprecipitation, chromatin extracts were added to the antibody-bound beads and incubated at 4°C overnight with rotation and then washed with Nexus washing buffer A to D (wash buffer A: 10 mM Tris-EDTA, 0.1% Triton X-100, wash buffer B: 150 mM NaCl, 20 mM Tris-HCl, pH 8.0, 5 mM EDTA, 5.2% sucrose, 1.0% Triton X-100, 0.2% SDS, wash buffer C: 250 mM NaCl, 5 mM Tris-HCl, pH 8.0, 25 mM HEPES, 0.5% Triton X-100, 0.05% sodium deoxycholate, 0.5 mM EDTA, wash buffer D: 250 mM LiCl, 0.5% IGEPAL CA-630, 10 mM Tris-HCl, pH 8.0, 0.5% sodium deoxycholate, 10 mM EDTA). End repair and dA-tailing were performed using the NEBNext End Repair Module and the NEBNext dA-Tailing Module. ChIP-nexus adaptors with mixed fixed barcodes (CTGA, TGAC, GACT, ACTG) were ligated with Quick T4 DNA ligase and converted to blunt ends with Klenow fragment and T4 DNA polymerase. The samples were treated with lambda exonuclease and RecJf exonuclease for generating Pol II footprints at high resolution. After each enzymatic reaction, the chromatin was washed with the Nexus washing buffers A to D and Tris buffer (10 mM Tris, pH 7.5, 8.0, or 9.5, depending on the next enzymatic step).

After RecJf exonuclease digestion, the chromatin was eluted and subjected to reverse crosslinking and ethanol precipitation. Purified single-stranded DNA was then circularized with CircLigase, annealed with oligonucleotides complementary to the BamHI restriction site and linearized by BamHI digestion. The linearized single-stranded DNA was purified by ethanol precipitation and subjected to PCR amplification with NEBNext High-Fidelity 2X PCR Master Mix and ChIP-nexus primers. The ChIP-nexus libraries were then gel-purified with QIAquick Gel Extraction Kit and quantified with Qubit™ dsDNA HS Assay Kit before sequencing with Illumina NextSeq 500.
**RNA-seq experiment**

For each RNA-seq experiment, total RNA levels were extracted from 107 kc167 cells using a RNeasy Mini Kit (Qiagen) and sent for library preparation using the TruSeq Stranded Total RNA Library Prep (Illumina). Reads were sequenced with an Illumina NextSeq 500 platform.

**Embryo collections**

Wild-type embryos (Oregon R) were collected from six population cages (28 × 17 × 17 cm, with 10,000–12,000 flies each, maintained in fly incubators at 25°C and 60% humidity) on 15 cm apple juice plates with yeast paste after pre-clearing. For ChIP-nexus experiments, the collection windows were 1–2 hr., 2–3 hr., 2–4 hr., 6–8 hr. after egg deposition (AED). Embryos were dechorionated with bleach and cross-linked with 1.8% formaldehyde in 2.5 ml Hepes buffer and 7.5 ml heptane, while vortexing at medium speed for 15 min. Embryos were devitellinized in methanol/heptane and kept at −20°C in methanol for up to 3 months until needed. Staining with DAPI was performed by standard methods and the embryos were analyzed using a fluorescence microscope to remove out-of-stage embryo contaminations.

5.3 **Quantification and statistical analysis**

All statistical analyses and plots were generated using the R program.

**TSS annotation of Drosophila melanogaster kc167 cells**

TSSs from non-overlapping flybase protein coding genes (fb-r5.47) were re-annotated using CAGE-seq (modENCODE_5333) data from Drosophila melanogaster kc167 cells. For TSS tag clustering, replicates were merged, and low-quality reads were removed with CAGER. Clusters with only one TSS (singletons) were discarded unless the normalized signal (TPM) was equal or above 5. TSSs within 30-bp of each other were clustered together and only clusters with at least 3 TPM from all TSS positions were
considered expressed and kept for further analysis. Varying the cutoffs changed the number of clusters and number of genes within promoter types, but did not change the main conclusions.

**Occupancy heatmaps**

In figure 2-2A all active promoters were sorted based on the decreasing levels of TAF1, calculated as the sum of the TAF1 ChIP-nexus signal (under control conditions) within a 101-bp window centered on the TSS. For each factor, the ChIP-nexus signal in RPM was calculated per base-pair for each individual promoter within a 1001-bp window around the TSS and plotted as a heatmap.

In figure 2-7B TBP heatmaps for each cluster were generated by plotting the normalized signal (between 0 and 1) per base-pair for each individual promoter within 101-bp window after aligning to the TSS.

**Correlation heatmap**

In figure 2-2B, Pearson correlation coefficients (PCC) were calculated by correlating the total ChIP-nexus signal for each factor at all active promoters using a 101-bp around the TSS. PCC values for each factor were transformed into distance measures by subtracting 1 and clustered hierarchically using the ward.D2 method.

**Average binding profiles**

For each factor, two ChIP-nexus replicates were merged after read count normalization. For all the metaprofiles, the average ChIP-nexus signal was calculated at the corresponding group of promoters in reads per million (RPM) after aligning to the TSS and a data smoothing of 5 was performed.

In figure 2-3A, the top 1,000 promoters with the highest TAF1 signal were selected by calculating the sum of the TAF1 ChIP-nexus signal (under control conditions) within a 101-bp window centered on the TSS. The position of footprints for each TFIID subunits was
defined as the middle point between the positive and negative strand summits (rounded if between bases).

**Focused vs Dispersed promoters**

To separate focused from dispersed promoters, the width of TCs was calculated by taking into the account the position and CAGE signal of TSSs along the cluster, which (Haberle et al. 2015). (1) Cumulative distribution of CAGE signal along the cluster is calculated. (2) A 0.1 and 0.9 were selected as lower (q1) and upper quantiles. (3) From the 5’ end, the position of each quantile is determined as the first base in which of the cumulative expression is higher or equal to the quantile percentage of the total CAGE signal of that cluster. (4) TC width is then calculated as the distance (in base pairs) between the two quantile positions. Promoters with a TC width \( \leq 9 \)-bp were considered to have a focused transcription initiation pattern while promoters with a TC width > 9 -bp were included in dispersed category.

**Selection of promoter types**

Promoter types were defined from the active promoter data set (\( n = 4300 \)) and based on the presence or absence of the motifs listed in table 4-1 (up to one mismatch was allowed for TATA and PB). TATA and TCT promoters were only required to contain the TATA-box and the TCT element, respectively, but were not excluded for having other elements. Conversely, all the other promoter types were filtered to not have a TATA-box.

In figures 2-5A and 2-6E DNA sequences for each promoter group were obtained from the dm6 genome and represented as heatmaps. The motifs logos for each group in figure 2-4C were generated using the R package ggseqlogo.
Table 5-14. Core promoter elements information

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
<th>Start window</th>
<th>End window</th>
</tr>
</thead>
<tbody>
<tr>
<td>TATA</td>
<td>TATAWAWR</td>
<td>-35</td>
<td>-25</td>
</tr>
<tr>
<td>Inr</td>
<td>TCAKTY</td>
<td>-5</td>
<td>5</td>
</tr>
<tr>
<td>DPE</td>
<td>GGWYV</td>
<td>27</td>
<td>32</td>
</tr>
<tr>
<td>MTE</td>
<td>CSARCSSA</td>
<td>10</td>
<td>30</td>
</tr>
<tr>
<td>PB</td>
<td>KCGRWCG</td>
<td>25</td>
<td>32</td>
</tr>
<tr>
<td>DRE</td>
<td>ATCGAT</td>
<td>-70</td>
<td>-10</td>
</tr>
<tr>
<td>Ohler1</td>
<td>YGGYCACACTR</td>
<td>-7</td>
<td>7</td>
</tr>
<tr>
<td>Ohler6</td>
<td>YRGATWTTY</td>
<td>-30</td>
<td>-20</td>
</tr>
<tr>
<td>Ohler7</td>
<td>YCAKCNCTA</td>
<td>-70</td>
<td>10</td>
</tr>
<tr>
<td>TCT</td>
<td>YYCTTYYY</td>
<td>-10</td>
<td>10</td>
</tr>
</tbody>
</table>

Total signal distributions (occupancy boxplots)

The total signal distributions for the boxplots shown in figures 2-4, 2-5, 2-6 and 2-7 were calculated using the log2+1 of the sum of the ChIP-nexus signal for each factor within a 101-bp window centered at the TSS. All p-values were calculated in R with a two-sample Wilcoxon test.

K-means clustering

K-means was performed in R using the top 1000 focused promoters with the highest TBP total signal and with the following parameters: centers=4, nstart=100 and x=30. Clusters were calculated using the TBP ChIP-nexus signal at each position (standardized to the z-score) within a 71-bp window centered at the TSS. The TBP chip-nexus samples used for clustering were not subject to any type of drug, including the vehicle DMSO.

Promoter element enrichments

Drosophila core promoter elements shown in Figure 2-4C were identified by the presence of known core promoter sequences in a specified window relative to the TSS (table 4-1). In some cases, up to one mismatch was allowed from the consensus. For each cluster
and promoter element, the enrichment was calculated by determining the ratio between the fraction of the promoter element in the cluster and the fraction of the same promoter element in the other three clusters. The significance of the observed frequencies was calculated with Fisher's exact test and corrected for multiple testing with the Benjamini–Hochberg method.

**Pol II pausing half-life values**

Pausing half-life measurements were taken from Shao et al. (2017) and reprocessed to match the dm6 genome version. Briefly, promoters were selected if (1) the total Pol II signal in both control and Flavopiridol treated condition was high (top 25%), (2) a typical footprint was observed for Pol II (distance between positive and negative strand peak < 80 bp), (3) the position of the Pol II footprint was less than 80 bp downstream of the TSS and. 2422 promoters fulfilled these criteria. For each promoter, the Pol II signal was calculated in a 41 bp window centered on the pausing position the midpoint between Pol II positive and negative summits). To calculate the half-life of paused Pol II at each promoter, the Pol II time course measurements were fitted into an exponential decay model using non-linear regression. Promoters with a paused Pol II half-life longer than 60 min were floored to 60 min to eliminate inflated values due to noise. The 2422 promoters were then ranked and divided into five quintiles.

TFIID occupancy was then compared between promoters with short Pol II pausing half-lives (q1, hf <= 6 min) and promoters with long Pol II pausing half-lives (q5, hf >= 27 min). The relative differences in occupancy between both groups were calculated by dividing the median of signal at q5 by the signal at q1 and transformed to log2.

**Changes in TFIID occupancy after TRI**

Differences in the total occupancy after triptolide was calculated at top 1000 promoters with highest Pol II occupancy by dividing the sum of the signal under triptolide (log2) by the sum of the signal under the dmso vehicle (log2) using a 101-bp window
centered at the TSS. Values below or above 1 reflect a decrease or increase after triptolide, respectively, while values of 1 reflect no changes.

**RNA expression levels**

For both RNA-seq and CAGE-seq data reads from two replicate experiments were aligned to the dm6 genome and merged after TPM normalization using the software Salmon and the R package CAGEr, respectively.

**Public data sets used for analysis**

All the protein structures, promoter-bound TFIID (6MZN), TFIIA-TBP (1NVP), TAF1-7 (4OY2), TAF4-12 (1H3O), TAF6-9 (6MZN) and TAF11-13 (6MZN), were downloaded from the RCSB Protein Data Bank and colored appropriately using UCSF CHIMERA60.

CAGE-seq data for the kc167 (modENCODE_5333) cells and 0-2 (modENCODE_5338), 2-4 (modENCODE_5339) and 6-8 (modENCODE_5341) hours fly embryos was downloaded from the modEncode project. The TBP, TFIIA, TFIIB, TFIIF and the Pol II ChIP-nexus data had been previously generated from our lab (GSE85741) and are publicly available at the Sequence Read Archive (SRA).
References & Bibliography

12. Baumann, D.G. and Gilmour, D.S. 2017. A sequence-specific core promoter-


73. Deng, W. and Roberts, S.G.E. 2005. A core promoter element downstream of the TATA box that is recognized by TFIIB. *Genes & Development* 19(20), pp. 2418–
2423.


ATP hydrolysis direct promoter clearance by RNA polymerase II. *Cell* 77(1), pp. 145–156.


TATA box sequence to rate-limiting steps in transcription initiation by RNA polymerase II. *Journal of Molecular Biology* 277(5), pp. 1015–1031.


Chemical Reviews 115(6), pp. 2255–2273.


318. Tao, Y., Guermah, M., Martinez, E., et al. 1997. Specific interactions and


353. Wong, K.H., Jin, Y. and Struhl, K. 2014. TFIH phosphorylation of the Pol


