Title
The punctilious RNA polymerase II core promoter

Permalink
https://escholarship.org/uc/item/83t0x888

Journal
GENES & DEVELOPMENT, 31(13)

ISSN
0890-9369

Authors
Ngoc, LV
Wang, Y-L
Kassavetis, GA
et al.

Publication Date
2017-07-01

DOI
10.1101/gad.303149.117

Peer reviewed
The Punctilious RNA Polymerase II Core Promoter

Long Vo ngoc, Yuan-Liang Wang, George A. Kassavetis, and James T. Kadonaga

[Key words: RNA polymerase II, Core Promoter, Core Promoter Elements, Sequence-specific Transcription Factors, TBP, TBP-related Factors, Chromatin]

Running Head: The Punctilious Pol II Core Promoter

Section of Molecular Biology, University of California, San Diego, La Jolla, CA 92093-0347, USA

1Corresponding author.
E-mail: jkadonaga@ucsd.edu
The signals that direct the initiation of transcription ultimately converge at the core promoter, which is the gateway to transcription. Here we provide an overview of the RNA polymerase II core promoter in bilateria. The core promoter is diverse in terms of its composition and function, yet it is also punctilious, as it acts with strict rules and precision. We additionally describe an expanded view of the core promoter that comprises the classical DNA sequence motifs, sequence-specific DNA-binding transcription factors, chromatin signals, and DNA structure. This model may eventually lead to a more unified conceptual understanding of the core promoter.
The RNA polymerase II (Pol II) transcription system is a key component in the expression of protein-coding genes as well as many non-coding genes in eukaryotes. The initiation of Pol II transcription is mediated by a stretch of DNA known as the core promoter (for reviews, see Smale and Kadonaga 2003; Goodrich and Tjian 2010; Lenhard et al. 2012; Kadonaga 2012; Roy and Singer 2015; Danino et al. 2015). The core promoter is sometimes referred to as the 'gateway to transcription', as the signals that lead to the initiation of transcription ultimately converge at the core promoter.

In the past, the core promoter was often thought to be a generic element – a stretch of DNA with a TATA box that functions universally at all genes. It then became apparent, however, that the TATA box is present in only a small fraction of metazoan core promoters and that there are no universal core promoter elements. Further studies revealed the diversity of the core promoter in terms of its composition as well as its function.

Moreover, it became apparent that the core promoter is punctilious – precise sequences at precise locations are essential for core promoter function. Some core promoter elements are involved in enhancer-core promoter specificity as well as in specific biological networks. In addition, there are intriguing connections between chromatin structure (including histone modifications) in the core promoter region and transcriptional activity.

Here, we discuss the initiation of transcription in bilateria (bilaterally-symmetric animals) from the perspective of the Pol II core promoter. Topics include the nature of transcription start sites (TSSs), core promoter sequence motifs, enhancer-promoter specificity, TATA box-binding protein (TBP) and related factors, transcriptional directionality, and an overall view of the components that contribute to the initiation of transcription. We will focus, in particular, on data derived from functional analyses of core promoter elements. We will not, however, include discussion of CpG islands, in which mammalian promoters are frequently located, but rather direct the reader to excellent review articles on this subject (Deaton and Bird 2011; Schübeler 2015). A few underlying themes in this essay are the diversity of core promoters, the punctilious
nature of core promoters, and the multifarious components that contribute to core promoter function. It is notable that many fundamental and important questions about the core promoter have yet to be answered.

**Focused vs. dispersed transcription patterns**

There are different transcription initiation patterns that are observed with Pol II (Fig. 1). Purified Pol II itself does not specifically recognize the core promoter. Instead, Pol II and a set of auxiliary factors (for instance, TFIIA, TFIIB, TFII D, TFIIE, TFIIF, and TFIIH at TATA box-dependent promoters) assemble into a transcription pre-initiation complex (PIC) at the core promoter. [Note: TFII D, which consists of TBP and ~13 to 15 TBP-associated factors (TAFs), is a key factor in the recognition of sequence motifs at the core promoter.] Upon addition of the ribonucleoside 5’-triphosphates, transcription rapidly initiates from the PIC (for a recent review, see Sainsbury et al. 2015). A pattern of transcription from a single site or a narrow cluster of sites (≤ 5 nt window) is probably derived from a specific PIC at a core promoter (see, for example, Kadonaga 1990). We refer to this TSS pattern as "focused" (also known as "narrow peak", "peaked", "sharp peak", and "single peak") (see also Juven-Gershon et al. 2008a; Kadonaga 2012).

In contrast to focused transcription, there is also "dispersed" (also known as "broad" and "weak") transcription, in which there is a pattern of several weak TSSs that are distributed over a region that might span 50-100 nt. The mechanisms and factors that are involved in dispersed transcription remain to be determined. Transcription from the multiple TSSs may occur via a mechanism that involves multiple PICs. It is also notable that dispersed promoter regions are deficient in ATG codons (termed "ATG deserts"; Lee et al. 2005a). The presence of an ATG desert would enable a single protein to be encoded by a promoter region with multiple TSSs.

Focused and dispersed TSS patterns represent two endpoints of a spectrum of transcription mechanisms, and mixed promoters, such as those with multiple weak TSSs and a
major predominant TSS, are often observed. The range of transcriptional patterns at promoters is sometimes referred to as "promoter shape".

Focused TSSs are frequently observed in regulated promoters, whereas dispersed TSSs are typically associated with ubiquitously expressed promoters (Hoskins et al. 2011). In addition, promoter shape is generally conserved between species (Carninci et al. 2006; Main et al. 2013). Moreover, the analysis of 81 different *Drosophila melanogaster* lines revealed that focused promoters are more evolutionarily constrained than dispersed promoters (Schor et al. 2017). From a teleological standpoint, it might be advantageous for regulated genes to be turned on and off at a single TSS at focused promoters and for constitutively active genes to maintain a steady stream of transcription via multiple TSSs at dispersed promoters.

It is also relevant to note that a key technical issue in the study of focused and dispersed promoters is the accurate determination of the TSSs. For example, processing or degradation of transcripts could lead to the inadvertent misidentification of TSSs. To minimize this problem, it is useful to map the 5' ends of capped nascent transcripts by using a method such as START-seq (Nechaev et al. 2010), GRO-cap (Kruesi et al. 2013), or 5'-GRO-seq (Lam et al. 2013). To date, however, most studies of promoter shape have been performed with accumulated steady-state RNAs. Hence, new insights might be gained from the analysis of promoter shape with TSSs that are determined by the mapping of nascent transcripts. For instance, recent analyses of nascent transcripts suggest that most human promoters have mixed (*i.e.*, combined focused and dispersed) transcription patterns (Lai and Pugh 2017) and that dispersed transcription occurs less frequently than previously thought from the analysis of steady-state RNAs (Scruggs et al. 2015; Core et al. 2014).

**Core promoter sequence motifs**

The activity of the core promoter is largely dependent upon the presence or absence of specific DNA sequences known as core promoter elements or motifs. Importantly, core promoters are
diverse not only in terms of the presence or absence of particular sequence motifs, but also with regard to the distinct functions that are mediated by specific core promoter elements. Some of the known core promoter motifs in bilaterians are shown in Fig. 2 and Table I. These sequence elements have been mostly studied in focused promoters.

There are no universal core promoter elements. Moreover, many core promoters lack any of the known motifs. Hence, there are probably other core promoter elements that remain to be discovered. Brief summaries of some core promoter motifs are as follows.

**The initiator (Inr)**

The initiator (Inr) motif is probably the most widely used core promoter motif in bilateria. It was originally found by Chambon and colleagues (Corden et al. 1980), and was incisively articulated as a discrete core promoter element by Smale and Baltimore (1989). The Inr encompasses the TSS, and is recognized by the TAF1 and TAF2 subunits of TFIID (Chalkley and Verrijzer 1999; Louder et al. 2016).

In human cells, the analysis of focused TSSs in nascent transcripts (5′-GRO-seq and GRO-cap methods) revealed the Inr consensus sequence of BBCA, BW (where, B is C/G/T and W is A/T) (Vo ngoc et al. 2017; for earlier versions of the Inr consensus, see Javehery et al. 1994; Lo and Smale 1996; Carninci et al. 2006). Over half of focused human promoters contain either a perfect match to the BBCA, BW Inr consensus or an Inr-like sequence with only a single mismatch outside of the CA, central core (Vo ngoc et al. 2017). To test the Inr consensus further, we analyzed focused TSSs in nascent transcripts (Start-seq method) from mouse cells (Scruggs et al. 2015), and also observed the precise placement of the same BBCA, BW Inr consensus in the core promoter (Fig. S1). This precision in the positioning of the Inr consensus sequence is an example of the punctilious nature of the core promoter.

The human and mouse BBCA, BW Inr consensus is similar but not identical to the *Drosophila* Inr consensus, TCA, GTY (where, Y is C/T) (Chalkley and Verrijzer 1999; Purnell et
al. 1994; Ohler et al. 2002; FitzGerald et al. 2006). The *Drosophila* Inr consensus appears to be a more restrictive version of the human/mouse Inr consensus. From an evolutionary perspective, it would be interesting to determine the Inr consensus sequences in diverse organisms and perhaps gain insight into whether the Inr had become more restrictive in *Drosophila* or less restrictive in mammals.

The A_{+1} in the Inr consensus sequence is usually the major site of transcription initiation and is designated as the +1 TSS position. The A_{+1} notation provides a specific reference point in the core promoter whether there is a single TSS or a cluster of TSSs. In addition, other core promoter motifs, such as the DPE and MTE (see below), function with the Inr and are strictly positioned with respect to the A_{+1} in the Inr consensus.

*The TATA box*

The TATA box is the first discovered core promoter motif in eukaryotes (Goldberg 1979) and was named after the TATAAAA sequence that is present in some upstream promoter regions. It is bound by the TBP subunit of the TFIID transcription factor (Sainsbury et al. 2015). The TATA box and TBP are ancient, as both are present in Archaea and eukaryotes (see, for example, Blombach et al. 2016).

The TATA box consensus has been investigated by the analysis of promoter sequences [e.g., STATAWAWR (simplified version of position-weight matrix from Bucher 1990); STATAWAAR (Ohler et al. 2002); STATAAA, TATAWRD (FitzGerald et al. 2006); TATAAR (Vo ngoc et al. 2017), where W is A/T, R is A/G, S is C/G, and D is A/G/T] as well as by the study of TBP binding to DNA [e.g., STATATAAGS (Wong and Bateman 1994); TATATAWR (Patikoglou et al. 1999)]. These TATA sequences mostly share the TATAWR motif, which is recommended as a general TATA consensus with the upstream T located at a position from -32 to -28 relative to the +1 TSS.
Promoters with a strict adherence to the TATA box consensus are somewhat rare. For instance, only about 3.5% of focused human promoters were found to have a perfect match to TATAAR (with the upstream T located from -33 to -28 relative to the +1 TSS; Vo ngoc et al. 2017). Moreover, only about 28% of focused human promoters were observed to have WWWW (an extremely loose TATA-like sequence) in the region from -33 to -23 relative to the +1 TSS. Hence, most promoters lack TATA or TATA-like sequences, and it is important to understand the DNA sequence elements and transcription factors that mediate TATA-less transcription.

It is also useful to note that TATA-containing promoters may or may not have Inr motifs. In fact, in human focused promoters, the occurrence of the TATA box is higher in the absence of the Inr or Inr-like sequences than in the presence of the Inr or Inr-like sequences (Vo ngoc et al. 2017). These observations suggest that some TATA boxes can drive transcription in the absence of an Inr. In other instances, the TATA and Inr can function synergistically for the recruitment of TFIID in a process that exhibits a strict spacing dependence between the two elements (Emami et al. 1997). It has also been found that transcription from TATA+Inr promoters is facilitated by high mobility group A1 (HMGA1) protein and Mediator (Xu et al. 2011).

The BRE\textsuperscript{u} and BRE\textsuperscript{d}

The basal transcription factor TFIIB binds to the TBP-TATA box complex to form a ternary complex in which TFIIB interacts with TBP as well as DNA flanking the TATA box (Sainsbury et al. 2015). The TFIIB-DNA contact sites that are upstream and downstream of the TATA box are known as the TFIIB recognition elements, BRE\textsuperscript{u} and BRE\textsuperscript{d} (Lagrange et al. 1998; Deng and Roberts 2005). Because the sequence-specific interaction of TFIIB with DNA is dependent upon the binding of TBP to the TATA box, the presence of a TATA box is required for a promoter to have functional BRE motifs.
The BRE is immediately upstream of the TATA box, and the G in its consensus sequence (SSRCGCC; Lagrange et al. 1998) appears to be the single most important nucleotide. The BRE (consensus sequence, RTDKKKK; Deng and Roberts 2005) is located immediately downstream of the 8 nt version of the TATA box (e.g., TATATAWR). It should be noted, however, that these consensus sequences have not been confirmed or revised with more recent data and methodology.

Like TBP and the TATA box, TFIIB and the BREs are present in Archaea and eukaryotes (see, for example, Blombach et al. 2016). Hence, the BRE is an ancient promoter element. Yet, the functions of the BRE motifs are not yet known. Depending on the promoter context, they have been found to have a positive or negative effect upon transcriptional activity (Lagrange et al. 1998; Evans et al. 2001; Deng and Roberts 2005). Intriguingly, the BRE was also found to suppress the ability of Caudal, a sequence-specific DNA-binding transcription factor, to activate transcription from TATA-dependent promoters (Juven-Gershon et al. 2008b). These findings indicate that further investigation of the BRE motif is likely to reveal interesting and important aspects of basal and regulated transcription.

The TCT motif (polypyrimidine initiator)

The TCT motif (also known as the polypyrimidine initiator; Perry 2005) is present in the core promoters of nearly all of the ribosomal protein genes in Drosophila and humans (Parry et al. 2010). This element encompasses the TSS and has the consensus of YYC+1TTTY in Drosophila (Parry et al. 2010) and YC+1TYTYY in humans (Parry et al. 2010; Vo ngoc et al. 2017), where transcription initiates at C+1 rather than at A+1 as is seen in Inr-containing promoters. The term "TCT" motif refers to the TCT trinucleotide that frequently encompasses the +1 TSS.

The TCT motif is a rare motif that is found only in ribosomal protein gene core promoters and a small number of other promoters, many of which are associated with genes that encode
proteins involved in translation (Parry et al. 2010). In humans, it is estimated that approximately 1% of focused core promoters contain a TCT motif (Vo ngoc et al. 2017). Hence, the TCT motif is an example of a core promoter motif that is rare but biologically important. The TCT motif regulates the network of ribosomal protein genes, and is thus the complement to the RNA polymerase I and RNA polymerase III transcription systems, which synthesize ribosomal RNAs.

The TCT motif is distinct from the Inr, but a single T to A substitution can convert a TCT motif into a functionally active Inr (Parry et al. 2010). These findings further reveal the punctilious nature of the core promoter. Precise sequences and precise positioning are essential features of core promoter function. In addition, TCT-dependent transcription in *Drosophila* involves the use of TBP-related factor 2 (TRF2) instead of the more commonly used TBP (Wang et al. 2014; discussed in more detail below).

*The downstream core promoter element, DPE*

The downstream core promoter element (DPE) functions cooperatively with the Inr for TFIID binding and transcriptional activity (Burke and Kadonaga 1996). The DPE is commonly found in *Drosophila* (~30% of core promoters) and appears to be rare in humans (Burke and Kadonaga 1997; Kutach and Kadonaga 2000). The *Drosophila* DPE consensus is RGWYV from +28 to +32 (or RGWYVT from +28 to +33) relative to the A_+1_ in the Inr (Kutach and Kadonaga 2000; Kutach 2000), and the human DPE consensus remains to be determined. There is a strict spacing requirement in the positioning of the DPE and Inr, as an increase or decrease of only a single nucleotide between the two elements results in a several-fold decrease in transcriptional activity as well as a reduction in the binding of TFIID (Kutach and Kadonaga 2000). This strict positioning requirement is another example of the punctilious nature of the core promoter.

As mentioned above, the DPE has rarely been found in human core promoters. This may be due to the scarcity of the DPE in humans and/or the lack of understanding of the human DPE consensus sequence. Functional human DPE motifs that resemble the *Drosophila* DPE have
been found in the human *IRF1*, *CALM2*, and *TAF7* gene promoters (Burke and Kadonaga 1997; Zhou and Chiang 2001, 2002; Duttke 2014). Notably, human transcription factors exhibit higher activity with wild-type DPE motifs than with mutant DPE motifs (with nucleotide substitutions or alteration of the Inr to DPE spacing) both in cells and in biochemical experiments (Burke and Kadonaga 1997; Zhou and Chiang 2001, 2002; Lewis et al. 2005; Juven-Gershon et al. 2006; Duttke 2014). These findings indicate that human transcription factors can recognize and function with the DPE. It was also found that DPE-specific transcription in humans involves Mediator, casein kinase II (CK2), and positive coactivator 4 (PC4) (Lewis et al. 2005). In the future, it will be important to analyze further the sequence consensus, Inr-DPE spacing, abundance, and transcription factor requirements of human DPE.

**The motif ten element, MTE**

The motif ten element (MTE) was identified as an overrepresented sequence in *Drosophila* core promoters (Ohler et al. 2002) and then found to be a TFIID binding site and a core promoter element that functions cooperatively with a precisely positioned Inr (Lim et al. 2004). The original consensus of the MTE was CSARCSSAAC from +18 to +27 relative to the A+1 in the Inr. A more detailed analysis revealed that there are three key contact points for the binding of TFIID to the downstream core promoter region and that the first and second contact points constitute the MTE and the second and third contact points constitute the DPE (Theisen et al. 2010). Hence, a tentative revised MTE consensus is CGANC from +18 to +22 and CGG from +27 to +29 (Table I). In the future, it will be important to gain a unified understanding of the TFIID-DNA interactions in the downstream core promoter region that support core promoter activity.

In this regard, the structure of human TFIID bound to a super core promoter that contains TATA, Inr, MTE, and DPE motifs (Juven-Gershon et al. 2006) revealed contacts of the TAF1 and TAF2 subunits of TFIID with the downstream core promoter region (Louder et al. 2016).
Consistent with these findings, TFIID-DNA photocrosslinking experiments with a reagent that extends from the DNA backbone phosphate detected the close proximity of the TAF1 subunit of TFIID with the downstream promoter (Kutach 2000). In contrast, photocrosslinking studies with a reagent that extends from the DNA major groove indicated the close proximity of TAF6 and TAF9 (but not TAF1 or TAF2) to the MTE and DPE sequences (Burke and Kadonaga 1997; Theisen et al. 2010). Moreover, TAF6-TAF9 complexes were found to interact with the DPE (Shao et al. 2005). These different TFIID-DNAcontacts may be due to different conformations of TFIID, as seen, for example, by Cianfrocco et al. (2013). It will also be important to determine the functions of these TAF-DNA contacts in the formation of the PIC.

Other core promoter elements

Some additional core promoter motifs include the following. The XCPE1 and XCPE2 motifs (X core promoter elements 1 and 2) were identified in the hepatitis B virus X gene promoter (Tokusumi et al. 2007; Anish et al. 2009). The downstream core element (DCE) was found in the human β-globin promoter (Lewis et al. 2000), and it comprises three subelements in the +6 to +34 region situated in close proximity to TAF1 (Lee et al. 2005b). Three downstream elements, termed GLE, DPE-L1, and DPE-L2, were found in the +4 to +32 region of MHC class I promoters (Lee et al. 2010). Another downstream motif, the DTIE (downstream transcription initiation element), was identified in the microRNA miR-22 promoter (Marbach-Bar et al. 2016). Furthermore, because there are many promoters with no known core promoter elements, there may be as-yet-undiscovered motifs with interesting and important biological functions.

Enhancer-core promoter specificity

In addition to their role in the basal transcription process, core promoter motifs, such as the DPE and TATA box, are involved in the regulation of gene expression by transcriptional enhancers (Fig. 3). For instance, when test enhancers were placed between divergently transcribed
promoters, the *Drosophila* AE1 and IAB5 enhancers were found to activate transcription preferentially from the TATA-dependent *even-skipped* promoter relative to the DPE-dependent *white* promoter (Ohtsuki et al. 1998; the *white* promoter was found to be DPE-dependent by Kutach et al. 2000). Moreover, in studies that directly compared the ability of enhancers to activate transcription from a TATA- or DPE-dependent core promoter in the same context, both DPE- as well as TATA-specific enhancers were observed (Butler and Kadonaga 2001).

Enhancer-core promoter specificity was also seen at the genome-wide level in the comparison of a developmental core promoter (a synthetic core promoter with TATA, Inr, MTE, and DPE motifs) with a housekeeping core promoter (the TCT motif-containing ribosomal protein S12 gene promoter) (Zabidi et al. 2015). Hence, these findings reveal that transcriptional enhancers can distinguish between different core promoters, and indicate that the specificity between enhancers and their cognate promoters can be achieved, at least in part, via core promoter motifs.

Enhancer-core promoter specificity was further examined in the context of the homeotic (Hox) gene network in *Drosophila* (Juven-Gershon et al. 2008b). Nearly all of the *Drosophila* Hox genes contain DPE-dependent core promoters, and Caudal, a sequence-specific DNA-binding transcription factor and key regulator of the Hox gene network, preferentially activates transcription from DPE-dependent promoters relative to some but not all TATA-dependent promoters (Juven-Gershon et al. 2008b; Shir-Shapira et al. 2015). In addition, the presence of the BRE\textsuperscript{u} motif suppresses the ability of Caudal to function in conjunction with the TATA box. These results show that the DPE is an important component of the Hox gene network and that Caudal, an important regulator of this network, can function as a DPE-specific activator. Moreover, the ability of the BRE\textsuperscript{u} to suppress Caudal activation of a TATA box-containing promoter suggests a role of TFIIB and the BRE\textsuperscript{u} in the regulation of the activity of sequence-specific transcription factors.
The DPE motif is also overrepresented in the core promoters of *Drosophila* genes that are regulated by Dorsal, a sequence-specific transcription factor that is a member of the NF-κB family of proteins (Zehavi et al. 2014). The DPE is essential for Dorsal-mediated activation of many genes that control dorsal-ventral patterning. In addition, in some promoter contexts, Dorsal preferentially activates transcription via the DPE relative to the TATA box.

How might transcription factors activate transcription preferentially via the DPE relative to the TATA box? It is known, for instance, that NC2 (negative cofactor 2; also known as Dr1-Drap1) as well as the Mot1 ATPase repress TATA-dependent transcription and activate DPE-dependent transcription (Willy et al. 2000; Hsu et al. 2008). It is therefore possible that DPE-specific activators can recruit factors such as NC2 and/or Mot1 to the core promoter and thus promote DPE-dependent transcription relative to TATA-dependent transcription. However, the mechanisms of core promoter motif-specific activation remain to be determined.

Lastly, it is relevant to mention that specificity for core promoter motifs applies not only to distant transcriptional enhancers, but also to promoter-proximal activator binding sites. For instance, an activating region that is 60 bp upstream of the mouse terminal deoxynucleotidyltransferase gene promoter exhibits a preference for the Inr relative to the TATA box (Garraway et al. 1996).

**TBP-related factors (TRFs)**

TBP and TFIIIB are present in Archaea and eukaryotes. Prior to the evolution of eukaryotes, it is likely that mechanism of transcription involved the binding of TBP to the TATA box and the subsequent assembly of TFIIIB, the RNA polymerase, and other factors into the pre-initiation complex (see, for example, Blombach et al. 2016). In this manner, the central role of TBP in the transcription process would have been established.

In bilateria, three additional TBP-related factors (TRFs) have been identified (see, for example: Goodrich and Tjian 2010; Akhtar and Veenstra 2011). These factors possess many of
the key features of TBP, such as sites of interaction with TFIIB and TFIIA, and therefore have much of the transcriptional potency of TBP. We refer to TBP and the TRFs as "system factors" (Duttke et al. 2014; Duttke 2015). TBP and the TRFs regulate gene expression via the basal transcription process. Some of the properties of the TRFs are discussed next.

**TBP-related factor 1 (TRF1)**

TRF1 (also known as TRF) was the first TBP-related factor to be identified (Crowley et al. 1993). TRF1 has been found only in insects. It can bind to the TATA box along with TFIIA and TFIIB and substitute for TBP in the transcription of some Pol II promoters in vitro (Hansen et al. 1997; Holmes and Tjian 2000). Moreover, TRF1 associates with BRF1 (an RNA polymerase III transcription factor) and mediates tRNA gene transcription from Pol III promoters (Takada et al. 2000; Isogai et al. 2007a; Verma et al. 2013). Thus, TRF1 participates in both Pol II and Pol III transcription. It is also interesting to note that the emergence of TRF1 did not appear to add any new transcriptional functions, but rather resulted in the subdivision and/or sharing of the pre-existing functions between TBP and TRF1.

**TBP-related factor 2 (TRF2)**

TRF2 (also known as TBPL1, TLP, TRP, and TLF) is present in bilateria (Duttke et al. 2014). Unlike TBP and the other TRFs, TRF2 does not bind to the TATA box and does not appear to possess any sequence-specific DNA-binding activity (Dantonel et al. 1999; Rabenstein et al. 1999; Wang et al. 2014). It does, however, interact with TFIIA and TFIIB (Rabenstein et al. 1999; Teichmann et al. 1999).

In *Drosophila*, TRF2 is involved in several different transcriptional programs. First, TRF2 associates with DREF (DNA replication-related element-binding factor) and activates transcription via the binding of DREF to DRE motifs (DNA replication-related elements) in promoters (Hochheimer et al. 2002). Second and third, by DRE-independent processes, TRF2,
but not TBP, is required for transcription from TCT-dependent as well as DPE-dependent core promoters (Hsu et al. 2008; Wang et al. 2014; Kedmi et al. 2014). The DPE functions with the Inr but not with the TCT element (Parry et al. 2010); hence, TRF2-driven transcription via the TCT motif probably occurs by a different mechanism than TRF2-mediated transcription via the DPE. Fourth, TRF2, but not TBP, is required for transcription of the histone H1 promoter by a process that does not appear to involve the DRE, TCT, or DPE (Isogai et al. 2007b). The partitioning of the transcriptional functions of TBP, TRF1, and TRF2 in Drosophila is depicted in Fig. 4.

The majority of the TRF2-dependent promoters in Drosophila lack a TATA box (see, for example, Isogai et al. 2007b; Wang et al. 2014). Given that TRF2 does not bind to the TATA box, these findings suggest that a key early function of TRF2 may have been to mediate TATA-less transcription (see, for instance, Duttke et al. 2014). It thus appears that, in contrast to the situation with TBP and TRF1 (see above), the combination of TBP and TRF2 has resulted in an expansion of the range of transcriptional mechanisms relative to those employed by TBP alone. This increase in the number of transcriptional programs led to the suggestion that the emergence of TRF2 facilitated the evolution of the bilateria (Duttke et al. 2014).

The loss of TRF2 is embryonic lethal in Drosophila (Kopytova et al. 2006), C. elegans (Kaltenbach et al. 2000; Dantonel et al. 2000), zebrafish (Müller et al. 2001), and Xenopus (Veenstra et al. 2000). In mice, however, TRF2 is not essential but is required for spermiogenesis (Martianov et al. 2001; Zhang et al. 2001; Zhou et al. 2013). The viability of TRF2-deficient mice could be due to the presence of a functionally analogous protein that can compensate for the absence of TRF2. It is also possible that the role of TRF2 changed substantially between frogs and mice. For instance, TBP or some other factor might have usurped the transcriptional function of TRF2 at most promoters and thus rendered TRF2 dispensable.
**TBP-related factor 3 (TRF3)**

TRF3 (also known as TBPL2 and TBP2) is found in vertebrates and is the TRF that is most closely related to TBP (Persengiev et al. 2003). It can bind to the TATA box, interact with TFIIA and TFIIIB, and mediate Pol II transcription in vitro (Bártfai et al., 2004; Jallow et al. 2004). TRF3 is present in a variety of mouse and human cell lines and tissues (Persengiev et al. 2003), but has particularly high expression in the testis and ovary in zebrafish (Bártfai et al. 2004) and *Xenopus* (Xiao et al. 2006) and in the ovary in mice (Xiao et al. 2006; Gazdag et al. 2007). TRF3 is required for normal embryonic development in zebrafish (Bártfai et al. 2004; Hart et al. 2007) and *Xenopus* (Jallow et al. 2004). In mice, however, the loss of TRF3 has no apparent phenotype except for female sterility, which is due to the requirement of TRF3 for the differentiation of female germ cells (Gazdag et al. 2009).

In zebrafish, TRF3 interacts with TAF3 (TBP-associated factor 3) and is essential for the expression of the *mespa* gene, which is required for hematopoiesis (Hart et al. 2007, 2009). Moreover, as seen with TRF3, the depletion of TAF3 also results in the failure to undergo hematopoiesis. In mouse cells, a complex that contains TRF3 and TAF3 was found to be involved in muscle cell differentiation (Deato and Tjian 2007; Deato et al. 2008). However, further studies, which included the analysis of TRF3 knockout mice, suggest that TBP, and not TRF3, remains active during muscle cell differentiation despite the rapid and dramatic loss of TBP protein in myotubes (Gazdag et al. 2009; Li et al. 2015; Malecova et al. 2016). It is possible that some of the reported differences could be due to the presence of an unknown factor that bypasses the need for TRF3 in muscle differentiation and acts in the organism but not in cells in culture.

**The core promoter is unidirectional**

In mammals, promoter regions frequently exhibit divergent transcription, with noncoding reverse-direction transcription that initiates upstream of the forward-direction TSS (Core et al.
Further analysis of this phenomenon led to a simple model in which core promoters are unidirectional and divergent promoter regions consist of forward- and reverse-direction core promoters (Duttke et al. 2015a; see also Andersson et al. 2015; Duttke et al. 2015b) (Fig. 5). The two opposing core promoters flank a central nucleosome-free region with binding sites for sequence-specific transcription factors. More generally, however, it is important to note that the analysis of the directionality of any particular promoter region should include the positions and orientations of all of the transcriptional elements, which include not only the core promoter motifs but also the binding sites for sequence-specific factors (see, for example, O'Shea-Greenfield and Smale 1992). In the future, it will be interesting to elucidate the biological functions of divergent transcription, such as a possible role in facilitating the evolution of new genes (Wu and Sharp 2013). In addition, transcription at one locus can increase transcription at a nearby locus (see, for instance, Engreitz et al. 2017); hence, reverse-direction transcription might enhance the level of forward-direction transcription.

An expanded view of the core promoter

Traditionally, the core promoter has been thought to comprise the TATA box, Inr, and other DNA sequence motifs that direct the assembly of the basal transcription machinery (i.e., Pol II, TFIID, TFIIB, and other auxiliary factors) into the PIC. It is now, however, useful to expand our perspective of the core promoter. Specifically, we could view the core promoter as a multidimensional element with some of the following components.

Role of sequence-specific DNA-binding transcription factors

Although it is well established that DNA recognition sites for the basal transcription machinery (e.g., TATA box, Inr, DPE) are important core promoter elements, it is also likely that binding sites for sequence-specific transcription factors (SSTFs), such as Sp1, can direct transcription
Vo ngoc et al. (Kadonaga) 19

initiation in conjunction with a motif such as the Inr. For example, a synthetic promoter that comprises a cluster of Sp1 binding sites and an Inr exhibits transcriptional activity that is similar to that of a TATA + Inr core promoter (Smale et al. 1990; Emami et al. 1995). It is thus reasonable to postulate that a SSTF recognition site (or sites) in the immediate upstream promoter region (~50-80 nt upstream of the TSS) could function with an Inr in lieu of a TATA box (Fig. 6A). Given the presence of the Inr or Inr-like sequences in over half of human focused promoters (Vo ngoc et al. 2017), SSTF binding site + Inr promoters may be widely used in mammals. It will therefore be important to examine this mechanism of transcription, particularly in the context of natural promoter regions.

Chromatin signals and structure

Transcription occurs in the context of chromatin, and hence, variations in the structure and composition of chromatin have considerable potential to influence the events that lead to transcription initiation. Multiple new lines of evidence are revealing intriguing connections between chromatin and transcription initiation (Fig. 6B). These findings suggest that chromatin signals and structure are components of an expanded version of the core promoter. Some examples are as follows.

First, there is an interesting connection between the TAF3 subunit of the TFIID complex and trimethylated histone H3K4 (H3K4me3), which is commonly found in the region immediately downstream of active promoters (Vermeulen et al. 2007; Lauberth et al. 2013). TAF3 binds to H3K4me3, and this interaction facilitates the assembly of the PIC. Thus, the TAF3-H3K4me3 interaction provides a means of recruiting TFIID to the core promoter region. In humans, H3K4me3 is present in only about 0.1% of the total histone H3 species (Young et al. 2009). Hence, H3K4 trimethylation can potentially add considerable specificity to the recruitment of TFIID to active promoters. In addition, the interaction of TFIID to promoters could be
augmented by histone acetylation, as the TAF1 subunit of TFIID contains a double bromodomain that can bind to diacetylated histone H4 (Jacobson et al. 2000).

Second, histone and DNA methylation have important roles in the promoter recruitment of RNA polymerases IV and V, which are specialized variants of Pol II in plants (for review, see, Haag and Pikaard 2011). Specifically, Pol IV is recruited to promoters containing methylated H3K9 via SHH1, a Pol IV-interacting protein that binds to unmethylated H3K4 and methylated H3K9 (Law et al. 2013; Zhang et al. 2013). Pol V is recruited to promoters containing methylated DNA via factors (DRD1 subunit of the DDR complex, SUVH2, SUVH9) that bind to the polymerase as well as to different forms of methylated DNA (Johnson et al. 2014; Liu et al. 2014). These findings show that histone methylation as well as DNA methylation can serve as chromatin-based signals for the recruitment of RNA polymerases. The positive effect of DNA methylation upon Pol V transcription is in contrast to the repressive effect of CpG methylation in vertebrates. Nevertheless, it is possible that DNA modifications such as methylation or hydroxymethylation could be used to recruit Pol II to promoters in animals.

Third, the prenucleosome, a stable conformational isomer of the nucleosome that associates with ~80 bp DNA, appears to be present in the "nucleosome-depleted region" (also known as "nucleosome-free region") that is located immediately upstream of the TSS of active promoters (Fei et al. 2015; Khuong et al. 2015). In the examination of the yeast PHO5 promoter in vivo, prenucleosome-like particles were observed at active promoters, whereas nucleosome-like particles were seen at repressed promoters (Brown et al. 2013; Fei et al. 2015). In addition, MPE-seq analysis with mouse embryonic stem cells revealed prenucleosome-like particles (i.e., histone-containing particles associated with ~61-100 bp DNA) in the immediate upstream region of active promoters but not inactive promoters (Ishii et al. 2015; Khuong et al. 2015). These findings suggest that prenucleosomes or prenucleosome-like particles are present in the nucleosome-depleted region of active promoters. Moreover, histone H3K56 can be acetylated by p300 in prenucleosomes but not in nucleosomes (Fei et al. 2015). It remains to be
determined, however, whether prenucleosomes participate in the transcription process. Notwithstanding, the association of prenucleosomes with only ~80 bp DNA suggests that they might be more permissive to transcription than canonical nucleosomes.

Fourth, histone variants might also participate in core promoter function. For example, H2A.Z- and H3.3-containing nucleosomes have been found at sites of active chromatin, such as promoters (Jin et al. 2009). Histones H2A.Z and H3.3 may destabilize nucleosomes (see, for example, Jin and Felsenfeld 2007) and thus facilitate transcription. In considering the presence of the human histone variants at promoters, it may be useful to note that H3.3 constitutes ~10% of the total histone H3 species and H2A.Z is about 1% to 3% of the total H2A species (Dang et al. 2016). Hence, the presence of H3.3 and H2A.Z at promoters could provide some specificity to core promoter function, but their roles, if any, in the initiation of transcription remain to be determined.

Thus, different aspects of the chromatin context are likely to be critical components of core promoter function. In such cases, however, it would be essential to understand the sources of the chromatin signals or structures that influence transcription.

Properties of DNA

Lastly, it seems likely that structural properties of DNA contribute to core promoter activity. For instance, the flexibility and curvature of DNA could facilitate interactions between transcription factors, and a decrease in the helical stability could increase the ability of Pol II to initiate transcription. However, an underlying DNA structure "code" for core promoters has yet to be determined. It is nevertheless interesting to note the general absence of core promoter DNA sequence motifs between the TATA box and Inr as well as between the Inr and MTE (Fig. 2). These regions may lack core promoter sequence elements, but they probably have a DNA structure that facilitates the transcription process.
Summary and perspectives

The core promoter is a rich and complex regulatory element. It is diverse in terms of its composition as well as its function. The core promoter is also punctilious – it acts unidirectionally with strict rules and precision. For instance, the change of a T nucleotide to an A can change a TRF2-driven TCT-dependent core promoter to a TBP-driven Inr-dependent core promoter. Moreover, specific core promoter elements can be associated with biological networks. The DPE is present in nearly all of the Hox gene promoters in Drosophila, and the TCT motif is present in nearly all of the ribosomal protein gene promoters in Drosophila and humans. In addition, some transcriptional enhancers exhibit a strong preference for specific core promoter elements.

We have also described an expanded view of the core promoter that comprises the classical DNA sequence motifs (such as the TATA box, Inr, DPE) along with promoter proximal SSTF binding sites, chromatin signals, and DNA structure. Each of these components might be important to varying degrees at any particular core promoter. Even though it appears to add complexity to our definition of the core promoter, the expanded model may result in a more unified and coherent conceptual understanding of the core promoter.

A few decades ago, with the discoveries of the TATA box and Inr, it seemed like we had a good understanding of the core promoter. We have since found, however, that the core promoter is a complex multidimensional regulatory element. We hope that, in the future, we might, once again, at least have the impression that we understand the punctilious RNA polymerase II core promoter.
Acknowledgments

We thank E. Peter Geiduschek, Robert Tjian, Tammy Juven-Gershon, Jack Cassidy, Jia Fei, and Sascha Duttke for critical reading of the manuscript. We are also very grateful to Nicolas Young for helpful information on histone modification abundances and to Craig Pikaard for insightful discussion of RNA polymerases IV and V. We also apologize to our colleagues for any unintended errors, oversights, or omissions. J.T.K. is the Amylin Chair in the Life Sciences. This work was supported by National Institutes of Health grants R35 GM118060 and R21 HG008781 to J.T.K.
References


Burke TW, Kadonaga JT. 1997. The downstream core promoter element, DPE, is conserved from *Drosophila* to humans and is recognized by TAFII60 of *Drosophila*. *Genes Dev* 11: 3020-3031.


Smale ST, Schmidt MC, Berk AJ, Baltimore D. 1990. Transcriptional activation by Sp1 as directed through TATA or initiator: specific requirement for mammalian transcription factor IID. *Proc Natl Acad Sci USA* **87**: 4509-4513.


Figure Legends

Figure 1. Focused, dispersed, and mixed transcription initiation patterns. In focused transcription, there is either a single predominant TSS or a narrow cluster of TSSs that probably derive from a single PIC. In dispersed transcription, there are multiple weak TSSs spread over a ∼50 to 100 bp region that likely emanate from multiple PICs. Focused and dispersed transcription patterns are two endpoints of a spectrum of possible mechanisms, and a variety of mixed TSS patterns are commonly observed. TSS patterns are also known as promoter shape.

Figure 2. A plethora of core promoter sequence motifs for RNA polymerase II. A typical core promoter might have zero to three of the indicated core promoter elements. The locations of the sequence motifs are roughly to scale. The consensus sequences are listed in Table I.

Figure 3. Enhancer-core promoter specificity. This diagram depicts transcriptional enhancers that function selectively with DPE-dependent or TATA-dependent core promoters (Butler and Kadonaga 2001; Juven-Gershon et al. 2008). Enhancer-core promoter specificity has also been observed with a developmental core promoter (with TATA, Inr, MTE, and DPE motifs) versus a housekeeping core promoter (with the TCT motif) (Zabidi et al. 2015). (Adapted from Butler and Kadonaga 2001.)

Figure 4. Transcriptional programs that are directed by TBP, TRF1, and TRF2 in Drosophila. This diagram shows the partitioning of transcriptional functions between TBP, TRF1, and TRF2 in Drosophila. It appears that each of these system factors is responsible for a set of transcriptional programs. As discussed in the text, humans lack TRF1 and contain TBP, TRF2, and TRF3. Moreover, in humans, the specific functions of factors such as TRF2 remain to be clarified. (Adapted from Duttke et al. 2014.)
**Figure 5.** A model for divergent transcription. In this model, a promoter region that exhibits divergent transcription contains a unidirectional forward core promoter and a unidirectional reverse core promoter that flank binding sites for sequence-specific transcription factors.

**Figure 6.** Potential functions of sequence-specific DNA-binding transcription factors and chromatin signals at the core promoter. (A) Postulated role of sequence-specific DNA-binding transcription factors (SSTFs) in core promoter function. In this model, SSTF binding sites in the immediate upstream promoter region (~50 to 80 bp upstream of the TSS) function in a manner that is analogous to a TATA box. Thus, the combination of an SSTF binding site and an Inr could act as a core promoter. (B) A composite of the potential role of chromatin signals and structure in core promoter function. It may be necessary to analyze the core promoter in the broader context of chromatin. Examples discussed in the text include the following. H3K4me3 has been found to recruit TFIID via its TAF3 subunit (Vermeulen et al. 2007; Lauberth et al. 2013). Salt-labile nucleosomes containing the histone variants H2A.Z and H3.3 have been found at active chromatin (Jin et al. 2009). The prenucleosome, a conformational isomer of the nucleosome that interacts with ~80 bp DNA, appears to be present in the immediate upstream region of active promoters (Fei et al. 2015; Khuong et al. 2015). In plants, RNA polymerase V is recruited to promoters via methylated DNA (Johnson et al. 2014; Liu et al. 2014). Although CpG methylation is generally repressive in vertebrates, DNA modifications such as methylation or hydroxymethylation may also function as a positive signal for the initiation of Pol II transcription.
Focused Transcription

Dispersed Transcription

Mixed Transcription
System Factors

- TBP
  - RNA Pol I
  - RNA Pol II - TATA
  - RNA Pol III - U6 snRNA
  - Other Programs?

- TRF1
  - RNA Pol III - tRNA
  - RNA Pol II - TATA?
  - Other Programs?

- TRF2
  - RNA Pol II - TCT
  - RNA Pol II - DPE
  - RNA Pol II - DREF/DRE
  - Other Programs - H1?
Coding Region

Sequence-specific Transcription Factors

Core Promoter (reverse)

Core Promoter (forward)

Coding Region
A  
Sequence-Specific Transcription Factor Binding Site(s)  

B  
Prenucleosome  Nucleosome  
MeN  H2A.Z  H3K4me3  H3.3
<table>
<thead>
<tr>
<th>Motif</th>
<th>Location</th>
<th>Consensus</th>
</tr>
</thead>
<tbody>
<tr>
<td>TATA box</td>
<td>Upstream T at -32 to -28</td>
<td>TATAWR</td>
</tr>
<tr>
<td>BRE\textsuperscript{u}</td>
<td>Upstream of TATA box</td>
<td>SSRCGCC</td>
</tr>
<tr>
<td>BRE\textsuperscript{d}</td>
<td>-23 to -17</td>
<td>RTDKKKK</td>
</tr>
<tr>
<td>Inr</td>
<td>-2 to +4</td>
<td>TCA\textsubscript{+1}GT (\textit{Drosophila})</td>
</tr>
<tr>
<td></td>
<td>-3 to +3</td>
<td>BBCA\textsubscript{+1}BW (Human)</td>
</tr>
<tr>
<td>TCT</td>
<td>-2 to +6</td>
<td>YCY\textsubscript{+1}TTYY (\textit{Drosophila})</td>
</tr>
<tr>
<td></td>
<td>-1 to +6</td>
<td>YCY\textsubscript{+1}TYYY (Human)</td>
</tr>
<tr>
<td>XCPE1</td>
<td>-8 to +2</td>
<td>DSGYGGGRAS\textsubscript{+1}M</td>
</tr>
<tr>
<td>XCPE2</td>
<td>-9 to +2</td>
<td>VCYCRTRCM\textsubscript{+1}Y</td>
</tr>
<tr>
<td>MTE</td>
<td>+18 to +22</td>
<td>CGANC</td>
</tr>
<tr>
<td></td>
<td>+27 to +29</td>
<td>CGG</td>
</tr>
<tr>
<td>DPE</td>
<td>+28 to +32</td>
<td>RGWYV</td>
</tr>
<tr>
<td>DCE</td>
<td>Box I: +6 to +11</td>
<td>CTTC</td>
</tr>
<tr>
<td></td>
<td>Box II: +16 to +21</td>
<td>CTGT</td>
</tr>
<tr>
<td></td>
<td>Box III: +30 to +34</td>
<td>AGC</td>
</tr>
<tr>
<td>DTIE</td>
<td>+23 to +31</td>
<td>GSGRDNHGG</td>
</tr>
</tbody>
</table>

Table I. Consensus sequences of some core promoter elements.

W: A or T
R: A or G
S: G or C
D: A or G or T (not C)
K: G or T
Y: C or T
B: C or G or T (not A)
M: A or C
V: A or C or G (not T)
N: A or C or G or T (any base)
H: A or C or T (not G)
SUPPLEMENTAL MATERIAL

The Punctilious RNA Polymerase II Core Promoter

Long Vo ngoc, Yuan-Liang Wang, George A. Kassavetis, and James T. Kadonaga
Supplemental Figure S1. The mouse Inr from Start-seq data is essentially the same as the human Inr from GRO-cap or 5'-GRO-seq data. Mouse Start-seq data from Scruggs et al. (2015) were analyzed as described in Vo ngoc et al. (2017). (A) The Inr is the most abundant overrepresented sequence near the TSS. Motif discovery analysis of the −5 to +6 region (relative to the +1 TSS) was performed with 17,728 focused TSSs. The prevalence (Coverage) and \( P \)-values of the top four sequence motifs is shown. The arrow indicates the position of the TSS. (B) Sequence logo of the mouse Inr at focused TSSs. The sequences of the 5854 TSS peaks with a perfect match to BBCA\(+1\)BW were used to generate the logo. (C) The BBCA\(+1\)BW Inr occurs frequently in focused promoters in mice. TSS peaks were divided into three groups: perfect match (Inr), one mismatch outside of the central CA\(+1\) (Inr-like), and all other sequences (Non-Inr). The number of TSSs in each group are shown in parentheses.