The RNA polymerase II core promoter — the gateway to transcription
Tamar Juven-Gershon, Jer-Yuan Hsu, Joshua WM Theisen and James T Kadonaga

The RNA polymerase II core promoter is generally defined to be the sequence that directs the initiation of transcription. This simple definition belies a diverse and complex transcriptional module. There are two major types of core promoters — focused and dispersed. Focused promoters contain either a single transcription start site or a distinct cluster of start sites over several nucleotides, whereas dispersed promoters contain several start sites over 50–100 nucleotides and are typically found in CpG islands in vertebrates. Focused promoters are more ancient and widespread throughout nature than dispersed promoters; however, in vertebrates, dispersed promoters are more common than focused promoters. In addition, core promoters may contain many different sequence motifs, such as the TATA box, BRE, Inr, MTE, DPE, DCE, and XCP1, that specify different mechanisms of transcription and responses to enhancers. Thus, the core promoter is a sophisticated gateway to transcription that determines which signals will lead to transcription initiation.

Address
Section of Molecular Biology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0347, USA

Corresponding author: Kadonaga, James T (jkadonaga@ucsd.edu)

Introduction
The RNA polymerase II core promoter comprises the sequences that direct the initiation of transcription (for reviews, see [1,2,3,4,5]). Thus, in principle, the core promoter could be as simple as a single motif that serves as a universal transcription start site, or as complex as a unique set of sequence instructions for each promoter. Historically, the former model has often been presumed to be true, but emerging data indicate that there is considerable diversity in core promoter structure and function.

The objective of this review is to provide an overview of current topics that relate to the core promoter, with a particular emphasis on sequence motifs in core promoters. In addition, we have annotated core promoter-related data in papers that were published in the past two years. It should further be noted that the properties of core promoters and their cognate factors are not likely to be strictly absolute; hence, the principles and ideas described in this essay should be taken only as current working models.

Focused versus dispersed core promoters
The vast majority of research on core promoters has been devoted to the study of focused core promoters (Figure 1). In focused core promoters (also referred to as single-peak, or SP, promoters), there is either a single transcription start site or a distinct cluster of start sites in a short region of several nucleotides. Most eukaryotic core promoters appear to be focused core promoters. In vertebrates, however, only about one-third or less of core promoters are focused core promoters; instead, the vast majority of genes appear to contain dispersed core promoters (also known as BR [broad distribution], MU [multimodal], or PB [broad with dominant peak] promoters), in which there are a number of transcription start sites distributed over a broad region that might typically range from 50 to 100 nucleotides (Figure 1). [Note that dispersed core promoters should not be confused with alternate promoters, which are distinct and sometimes differentially regulated promoters that are typically located hundreds or thousands of nucleotides apart.]

Core promoter elements such as the TATA box, BRE, Inr, MTE, DPE, and DCE (Figure 2; discussed in greater detail below) are typically found in focused core promoters. These core promoter elements are not universal; rather, each is present in only a subset of core promoters. Moreover, some core promoters appear to lack all of the known core promoter elements. It is interesting to note that the TATA box and BRE are the most ancient of the core promoter motifs. The TATA box and BRE along with their cognate protein factors, TBP (TATA box-binding protein) and TFIIB (transcription factor IIB), are conserved from Archaea to humans (for review, see [6]). The TATA box is also present in plant promoters [7,8]. The MTE and DPE appear to be conserved among metazoans. By contrast, dispersed core promoters are typically found in CpG islands in vertebrates and gener-
ally lack TATA, DPE, and MTE motifs (see, e.g. [1,3,5,9,10]).

Thus, focused core promoters are more ancient and used in a much broader range of organisms than dispersed promoters. In addition, several of the key sequence motifs that contribute to the activity of focused core promoters have been identified. On the contrary, in vertebrates, dispersed core promoters are more common than focused promoters. Moreover, little is known about the sequences and factors that are responsible for transcription from dispersed core promoters. It is interesting to note, however, that the promoter region of dispersed TATA-less promoters are generally deficient in ATG triplets [11]. There may be fundamental differences in the basic mechanisms of transcription from focused versus dispersed core promoters.

The initiator (Inr)
The initiator (Inr) motif encompasses the transcription start site [1,12]. Based on functional assays, the Inr consensus was determined to be YYANWYY in humans and TCAKTY in Drosophila (degenerate nucleotides are indicated according to the IUPAC nucleotide code). The A nucleotide in the middle of the Inr consensus is often the +1 start site in focused core promoters. Inr-like sequences have also been described in Saccharomyces cerevisiae (e.g., see [13*] and references therein). The Inr is probably the most commonly occurring sequence motif in focused core promoters (see, e.g. [14,15*,16*]). The Inr is a recognition site for the binding of TFII D. Although a number of proteins have been found to bind to Inr sequences, the binding of TFII D to the Inr appears to be particularly important because the sequence specificity of TFII D binding to the Inr region of the core promoter is identical to the Inr consensus sequence [17].

The computational analysis of thousands of mammalian transcription start sites suggests that the mammalian Inr consensus is YR, where R corresponds to the +1 start site [10*,18*]. By contrast, the computational analysis of thousands of Drosophila core promoters reveals a much more strict consensus sequence of TCAGTY [14,15*]. This sharp difference in the specificity of the Inr consensus between Drosophila and mammals suggests that mammalian transcription factors have evolved to function with a broader range of Inr sequences than Drosophila transcription factors. This property may be related to the prevalence of dispersed core promoters in mammals but not in Drosophila.

The TATA box and BRE
The TATA box, which is the most ancient and most widely used core promoter motif throughout nature, was aptly the first eukaryotic core promoter element to be identified (Michael L Goldberg, PhD thesis, Stanford University, 1979). The TATA box has a consensus of TATAWAAR, where the upstream T nucleotide is most commonly at −31 or −30 relative to the A + 1 (or G + 1) in the Inr (see, for instance [10*,19*]). The TATA box is recognized and bound by TBP, which is a subunit of the TFII D complex in eukaryotes. As noted above, the TATA box is present in a subset of focused core promoters. Consequently, the TATA box appears to be
uncommon in vertebrates (see, e.g. [10,20,21]), because only about one-third or less of vertebrate core promoters is focused and only a fraction of the focused core promoters contains a TATA box.

Yet, in this context, it is useful to note parenthetically that it is probably a good practice to be somewhat cautious in the interpretation of expected frequencies of core promoter motifs. For instance, the frequency of occurrence of the TATA box is an estimate that depends on the parameters (i.e. sequences and positions) that are used to define the TATA box as well as the accuracy and promoter coverage of the particular database that is used in the analysis. Thus, although it is clearly apparent that there are many TATA-less promoters, the contribution of the TATA box or functionally equivalent sequences to vertebrate transcription remains to be determined unambiguously, and may be greater than is currently believed. Ultimately, it will be necessary to determine the actual function of each potential motif in each core promoter.

The BRE (TFIIB recognition element) was originally identified as a TFIIB-binding sequence that is immediately upstream of a subset of TATA boxes [22]. It was then found that TFIIB can bind upstream or downstream of the TATA box at the BREu (upstream BRE, which is the same as the original BRE) or BREd (downstream BRE) sequences [23,24]. The BREu consensus is SSRCGCC [22]. BREd is located immediately downstream of the TATA box and has a consensus of RTDKKKK [23]. Depending on the promoter context, the BREu and BREd can act in either a positive or negative manner [22,23,24].

The DPE and MTE
The DPE (downstream core promoter element) was identified as a downstream TFIID recognition sequence that is important for basal transcription activity [25]. The DPE is conserved from Drosophila to humans, and is located from +28 to +33 relative to the A +1 in the Inr. The DPE consensus is RGWYVT in Drosophila [26]. The DPE consensus in humans has yet to be determined; however, mammalian core promoters containing sequences that conform to the Drosophila consensus have been found to possess DPE activity. The DPE functions cooperatively with the Inr, and the spacing between the Inr and DPE is critical for optimal transcription. Photocrosslinking studies revealed that the DPE is in proximity to the TFIID subunits TAF6 (TAFI60) and TAF9 (TAFI40), which contain histone fold motifs and are related to histones H4 and H3 [27]. It is thus possible that the TAF6–TAF9 subunits of TFIID interact with the DPE in a manner that is similar to binding of histones H3–H4 to DNA in nucleosomes [28].

The MTE (motif ten element) was found by a combination of computational and biochemical studies [14,29]. The MTE has a consensus of CSARCSSAAC from +18 to +27 relative to A +1 in the Inr in Drosophila. Mutation of the nucleotides from +18 to +22 can abolish MTE activity in vitro and in cultured cells. Like the DPE, the MTE functions cooperatively with the Inr with a strict Inr-MTE spacing requirement. The addition of an MTE can compensate for the loss of basal transcription activity that occurs upon the mutation of a TATA box or a DPE. Moreover, the MTE exhibits synergy with the TATA and DPE motifs. This synergy between the MTE and other core promoter motifs inspired the design of a Super Core Promoter, SCP, which contains optimized versions of the TATA box, Inr, MTE, and DPE [30]. The SCP is the strongest known core promoters in vitro and in cultured cells. In addition, the SCP exhibits unusually high affinity for the binding of TFIID.

The available evidence indicates that the MTE is present in humans. First, mutation of the MTE in a Drosophila core promoter causes a reduction in transcription by human factors both in vitro and in cultured cells [29,30]. Hence, the human transcriptional machinery recognizes the MTE. Second, a human core promoter containing a functional MTE has been identified [29]. However, the MTE generally does not emerge as an overrepresented sequence in computational analyses of mammalian promoter databases (see, e.g. [15,18]). As described above for the Inr, it is possible that the MTE as well as the DPE may have broader and less restrictive consensus sequences in mammals than in Drosophila. In this regard, it is interesting to note that the analysis of mammalian core promoter sequences [18] revealed a ‘gcg motif’ (which may be identical to ‘motif8’ of [31]) at +20 as well as a ‘gcg echo motif’ at +30. The gcg motif overlaps with the mutationally sensitive +18 to +22 region of the MTE [29], and the gcg echo motif overlaps with the location of the DPE. Thus, the gcg motif may be the mammalian version of the MTE, whereas the gcg echo motif may correspond to the DPE.

The DCE and XCPE1 motifs
The DCE (downstream core element) was originally found in the human beta-globin promoter [32], and has also been characterized in the adenovirus major late promoter [33]. The DCE occurs frequently with the TATA box, and appears to be distinct from the DPE. The DCE consists of three subelements: S8, CTTC from +6 to +11; S11, CTGT from +16 to +21; and S11, AGC from +30 to +34. Photocrosslinking studies revealed that the DCE is in proximity to TAF1.

The XCPE1 (X core promoter element 1) motif is located from –8 to +2 relative to the +1 start site and has the consensus sequence of DSGYGRASM [34]. It is present in about 1% of human core promoters, most of which are TATA-less. XCPE1 exhibits little activity by itself. Instead, it acts in conjunction with sequence-specific
activators, such as NRF1, NF-1, and Sp1. Thus, XCPE1 may be a member of a larger family of motifs that work along with sequence-specific activators in CpG islands to direct transcription initiation.

In the future, it will also be important to investigate functional interactions between different core promoter motifs. Along these lines, computational studies have revealed the co-occurrence of various combinations of core promoter motifs [15,16,35-37]. These studies not only confirm previously known interactions between core promoter motifs but also identify new potential interactions.

**Diversity in core promoter function**

The existence of different core promoter elements results in diversity in core promoter function (reviewed in [38]). For instance, enhancers are functionally linked to core promoters (see, e.g. [39]), and some transcriptional enhancers have been found to exhibit specificity for TATA versus DPE core promoter motifs [40]. In addition, different factors mediate the basic transcription process from different types of core promoters. For example, a set of purified transcription factors (TFIIF, TFIIB, TFIID, TFIIH, RNA polymerase II, PC4, and Sp1) that are sufficient to transcribe a TATA-dependent core promoter were found to be unable to transcribe a DPE-dependent core promoter [41]. Moreover, in transcription reactions performed with a crude nuclear extract, it was found that NC2 (also known as Dr1-Drap1) stimulates DPE-dependent transcription and inhibits TATA-dependent transcription [42]. However, the enhancement of DPE-dependent transcription by NC2 was not seen in a purified reconstituted system [41], perhaps because of the absence of an additional auxiliary factor that was present in the crude extract but not in the purified system. In separate work, it was also found that the Inr element contributes to resistance of repression of TATA-dependent transcription by NC2 [43]. Furthermore, in RNAi depletion studies, TAF1 and TAF4 were observed to be important for transcription from a DPE-dependent reporter gene but not from a TATA-dependent reporter gene [44]. At the present time, we have an intriguing yet incomplete picture of the factors involved in transcription from different types of core promoters. This subject is an important area of future investigation.

**TBP-related factors**

Diversity in the function of the transcription machinery can be seen with TBP and TBP-related factors (TRFs) (for recent reviews, see [45,46,47]). There are three TRFs: TRF1, TRF2 (also known as TLF, TLP, TRF, and TRP), and TRF3 (also known as TBP2). TRF1 is absent in vertebrates but is present in *Drosophila*, in which it binds to a TC-rich sequence and mediates transcription by RNA polymerases II and III [48,49]. TRF2 does not bind to TATA box sequences, and is involved in RNA polymerase II transcription in *Drosophila* and vertebrates. *Drosophila* TRF2 is in a multisubunit complex that contains ISWI and DREF proteins [50]. TRF2 in *Drosophila* has also been found to exist in a short form and a long form, both of which associate with ISWI [51]. TRF3 is present in vertebrates but not in *Drosophila*, and is the TRF that is most closely related to TBP. TRF3 binds to TATA box sequences and participates in transcription by RNA polymerase II.

It is becoming further apparent that TBP, which had been previously thought to be a universal transcription factor, is not required for the transcription of many genes (see, for instance [52,53]). Instead, the available data indicate that differential functions of TBP and the TRFs are involved in many different regulatory networks. For example, TRF2 is required for the transcription of the TATA-less histone H1 gene but not the TATA-containing S-phase regulated core histone genes, whereas TBP exhibits the opposite effect on those genes [54]. In addition, opposite effects of TRF2 and TBP were seen with the TATA-less neurofibromatosis type 1 (NF1) promoter and the TATA-containing c-fos promoter [55]. Strikingly, a novel complex containing TRF3 and TAF3 replaces the TBP-containing TFIID complex during the differentiation of myoblast cells into myotubes [56]. These findings provide an example of a TBP to TRF switch during cellular differentiation. In this regard, it is interesting to note that TBP and TRF3 are expressed differently during mouse oogenesis [57]. TRF3 has also been found to be important for hematopoiesis in zebrafish [58].

**Conclusions and future prospects**

The core promoter is diverse and complex. We still need to gain a better understanding of the DNA sequences that dictate core promoter function as well as the protein factors that function at the different types of core promoters. It will be particularly important to devote more effort to the study of the mechanisms of transcription at dispersed core promoters, because current evidence (see, e.g. [11,34,59]) suggests that there may be fundamental differences in the strategies and mechanisms of transcription from dispersed versus focused core promoters.

It will be interesting to investigate whether the binding of RNA polymerase II to the Mediator complex (see, e.g. [60,61]) influences basal transcription activity. For instance, Mediator has been found to contribute to the transcription of a DPE-containing promoter [41]. Moreover, a new form of RNA polymerase II, termed Pol II(G), contains an additional protein named Gdown1 that enables RNA polymerase II to become responsive to Mediator [62].
From a broader perspective, it will also be important to understand how core promoters function in biological processes, such as gene networks and development. The acquisition and integration of this information will provide us with the knowledge of the events that revolve about the core promoter — the gateway to transcription.

Acknowledgments
We are grateful to Barbara Rattner, Timur Yusufzai, and Alexandra Lusser of core promoter function, as it has been found that transcription initiation complexes are formed with the assistance of the binding of the TAF3 subunit of TFIIID to H3K4me3-containing nucleosomes [70].

In addition, chromatin structure may be a component of core promoter function, as it has been found that transcription start sites are often flanked by the H2A.Z histone variant (see, e.g. [63,64,65,66,67]) and are frequently located immediately upstream of trimethylated histone H3 lysine 4 (H3K4me3) (see, e.g. [66,67,68,69]). It is possible, for instance, that transcription initiation complexes are formed with the assistance of the binding of the TAF3 subunit of TFIIID to H3K4me3-containing nucleosomes [70].

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
- of outstanding interest

17. Statistical analysis of Drosophila core promoter sequences led to predictions regarding the frequency of occurrence of the TATA (16%), Inr (88%), DPE (22%), and MTE (10%) motifs as well as specific combinations of core promoter motifs.
20. Development of a model for the prediction of transcription start site usage in mammals. This model is based on over-represented sequences that flank transcription start sites. Key sequence motifs include an Sp1-like recognition site, a TATA-box-like sequence, an Inr (Py.Pu at +1), and two downstream sequences termed ‘gcg motif’ and ‘gcg echo motif’. The latter two motifs bear a resemblance to the downstream MTE and DPE motifs, and may function as mammalian versions of the MTE and DPE to promote the binding of TFIIID. In addition, the GCG and GCG echo motifs may be related to ‘motif8 described in [31].’
22. Extensive analysis of mouse CAGE tags revealed a distinct preference for the location of the TATA box at –31 or –30 relative to the +1 transcription start site.
25. High-throughput transient transfection studies yielded 387 functional human promoters. About 16% of these functional promoters contained...
TATA box motifs. Deletion analysis of 45 promoters revealed that sequences from about –50 to –300 relative to the transcription start site generally activate transcription. Negative elements were found to be located from –350 to –1000 relative to the start site.


31. Creation of super core promoters (SCPs) for RNA polymerase II transcription by combining the TATA, Inr, MTE, and DPE motifs in a single promoter. The SCPs exhibit strong transcription in vitro and in vivo and have unusually high affinity for TFII D. This work also shows that the MTE and DPE motifs each contribute to core promoter activity in vivo in human cells.


33. Identified the recognition motif of YY1, a sequence-specific DNA-binding protein, as a commonly occurring sequence in the downstream region of promoters that contain short 5’ untranslated regions. In most instances, the YY1 recognition motif, which contains an ATG sequence, correlated with the expected translation start site. It was suggested that the YY1 recognition motif is involved in both transcription and translation. The promoter sequence analysis also revealed a GC-rich downstream promoter motif that was termed ‘motif8’. Motif8 may be related to the ‘gcg motif’ of [18].


37. Identification of a new core promoter element termed XCP E1, which is located from –8 to +2 relative to the +1 transcription start site. XCP E1 exhibits little promoter activity by itself; instead, it works in conjunction with sequence-specific activators such as NRF1, NF-1, and Sp1. It was estimated that about 1% of human core promoters contain an XCP E1 motif.


40. Analysis of core promoter sequences that are conserved between humans and mice suggests the use of TATA-Inr-MTE and BRE-Inr-MTE combinations.


42. Computational analysis of the co-occurrence of Drosophila core promoter motifs revealed distinct core promoter modules.


48. RNAi depletion analysis of TFIID subunits revealed a critical role of TAF4 in maintaining the integrity of the TFIID complex. TAF1 and TAF4, but not TAF5, were found to be important for the transcription of a reporter gene that contains a DPE motif. By contrast, transcription from a related TATA-dependent reporter gene was not affected by the depletion of TAF1 or TAF4.


Identified a large 175 kDa form of TRF2. The C-terminal region of the 175 kDa form of TRF2 is identical to the previously described 75 kDa TRF2 polypeptide. The two forms of TRF2 exist in Drosophila, are encoded by the same gene, and appear to have related functions.

Depletion of TRF3 in zebrafish embryos results in a defect in hematopoiesis. A key target of TRF3 was found to be the miR-430 microRNA. TBP appears to be involved in the degradation of maternal mRNA by the mammalian Mediator complex.


