Accepted Manuscript

The core promoter: at the heart of gene expression

Yehuda M. Danino, Dan Even, Diana Ideses, Tamar Juven-Gershon

PII: S1874-9399(15)00090-5
DOI: doi: 10.1016/j.bbagrm.2015.04.003
Reference: BBAGRM 874

To appear in: BBA - Gene Regulatory Mechanisms

Received date: 17 February 2015
Revised date: 19 April 2015
Accepted date: 23 April 2015

Please cite this article as: Yehuda M. Danino, Dan Even, Diana Ideses, Tamar Juven-Gershon, The core promoter: at the heart of gene expression, BBA - Gene Regulatory Mechanisms (2015), doi: 10.1016/j.bbagrm.2015.04.003

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.
The core promoter: at the heart of gene expression

Yehuda M. Danino¹, Dan Even¹, Diana Ideses¹ and Tamar Juven-Gershon¹*

¹The Mina and Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat Gan 5290002, Israel

Running title: The core promoter: a central player in gene expression

Key words: core promoter; RNA Pol II transcription; core promoter elements/motifs; enhancer-promoter specificity; core promoter preferential activation; gene expression

The authors declare that there are no potential conflicts of interest.

* To whom correspondence should be addressed. Tel: +972-3-531-8244; Fax: +972-3-738-4058; Email: tamar.gershon@biu.ac.il
ABSTRACT

The identities of different cells and tissues in multicellular organisms are determined by tightly controlled transcriptional programs that enable accurate gene expression. The mechanisms that regulate gene expression comprise diverse multiplayer molecular circuits of multiple dedicated components. The RNA polymerase II (Pol II) core promoter establishes the center of this spatiotemporally orchestrated molecular machine. Here, we discuss transcription initiation, diversity in core promoter composition, interactions of the basal transcription machinery with the core promoter, enhancer-promoter specificity, core promoter-preferential activation, enhancer RNAs, Pol II pausing, transcription termination, Pol II recycling and translation. We further discuss recent findings indicating that promoters and enhancers share similar features and may not substantially differ from each other, as previously assumed. Taken together, we review a broad spectrum of studies that highlight the importance of the core promoter and its pivotal role in the regulation of metazoan gene expression and suggest future research directions and challenges.
Introduction

Appropriate temporal and spatial gene expression is a highly complex process underlying the fate and function of different cells and tissues. The regulation of this process is composed of multiple levels and orchestrated molecular events [1-3]. A central event in the regulation of eukaryotic gene expression is the initiation of transcription. The initiation of transcription of protein-coding genes and distinct non-coding RNAs occurs following the recruitment of RNA polymerase II (Pol II) to the core promoter region by the basal transcription machinery [4].

The core promoter is generally defined as the minimal DNA sequence that directs accurate initiation of transcription. The core promoter sequence encompasses the transcription start site (TSS), typically referred to as the +1 position [5, 6]. Examination of the distribution of TSSs reveals that there are multiple modes of transcription initiation (Fig. 1A). Distinct molecular players can open the chromatin structure at the core promoter region and thus facilitate initiation of transcription. Interestingly, active promoters are associated with specific chromatin signatures. These include: nucleosome-depleted regions (NDR) or reduced nucleosome occupancy over the promoters, DNaseI hypersensitive sites (DHS) and the enrichment of specific histone modifications, such as di- and tri-methylation of H3K4 and acetylation of H3K4 and H3K27 (Fig. 1B) [7, 8]. Notably, both human and yeast nucleosomes that are upstream and downstream of the TSSs of multiple genes have been shown to correlate with the occupancy of the histone variant H2A.Z (termed Htz1 in yeast) (reviewed in [9]).

In the past, it was assumed that the core promoter is a generic entity that functions in a universal manner. Nowadays however, the growing convention is that the unique
properties of a given promoter are a function of its architecture and core promoter motifs composition (Fig. 1C and D) [5, 6, 10, 11].

The core promoter, which is often referred to as “the gateway to transcription”, is a central component in the initiation of transcription [12, 13]. Research in the past decade has enhanced our understanding of the fundamental roles that the core promoter plays in the initiation of transcription, as well as in the regulation of additional aspects of gene expression. Insights are gained from studies of specific genes and gene networks [13-15], as well as from genome-wide studies [11, 16] utilizing methodologies such as PEAT [17], 5' RACE [18], CAGE [19], FAIRE-seq [20], ChIP-seq [21], Gro-seq [22], and RNA-seq [23], and key projects and consortia (e.g. modENCODE [24], ENCODE [25] and FANTOM5 [26]), which developed following the implementation of some of the above methods. Accordingly, core promoters can be studied at different resolutions: from genomic architecture, transcription co-regulators and sequence-specific transcription factors (Fig. 2A), through basal transcription factors (Fig. 2B and C) and DNA sequence motifs (Fig. 2C). Importantly, the different experimental strategies complement each other and together, provide the elaborate view of core promoters. Here, we review the current state of knowledge relevant to the contribution of the core promoter to multiple aspects of gene expression, and discuss future directions and challenges in the field.

1. Diversity in the transcription initiation landscape

1.1. Multiple modes of transcription initiation

The core promoter is best known for its role in directing proper transcription initiation at the TSS. Two modes of transcription initiation, focused and dispersed, were
previously noted in metazoan (Fig. 1A) (reviewed in [6, 11]). Focused (also termed “sharp peak”) promoters contain a single predominant TSS or a few TSSs within a narrow region of several nucleotides [10]. Focused promoters encompass approximately between -40 to +40 nucleotides relative to the TSS (referred to as the +1 position). Focused transcription initiation is associated with spatiotemporally regulated tissue specific genes [27] and with canonical core promoter elements that have a positional bias, such as the TATA box, Initiator, MTE and DPE [28] (Fig. 1C).

Dispersed (also termed “broad”) promoters contain multiple weak start sites that spread over 50 to 100 nucleotides at the promoter region ([10, 11] and refs therein). Dispersed transcription initiation is associated with constitutive or housekeeping genes. Vertebrate dispersed promoters often contain CpG islands and Sp1 and NF-Y sites [6, 10, 29] whereas Drosophila core promoters often contain elements that have weaker positional biases (as compared to the focused promoters), but frequently co-occur in a specific order and orientation: Ohler 1, DNA replication element (DRE), Ohler 6 and Ohler 7 [28, 30] (Fig. 1D). Although the focused promoter architecture exists in all the organisms and is the predominant initiation mode in simpler organisms, the dispersed mode is more common in higher eukaryotes [10, 27]. For example, over 70% of vertebrate promoters are dispersed [29, 31-33]. From a teleological standpoint, the associations of sharp TSSs with regulated genes and of broad TSSs patterns with constitutively expressed genes are rather intuitive. It would be easier to achieve a more precise control of gene expression from focused TSSs, as compared with dispersed promoters of housekeeping genes, which would be constitutively transcribed with minimal variation of gene expression by usage of multiple start sites [10].

1.2. Focused versus Dispersed initiation patterns - recent studies, new insights
Despite the abovementioned distinction between the two modes of transcription initiation, classification of transcription initiation landscapes is not so straightforward. Functional experiments and genome-wide studies using advanced technologies imply that there are multiple ways to classify promoters. Thus, the boundaries between these two major types of promoters are sometimes unclear [6, 34]. With respect to the “focused vs. dispersed” sub-classifications mentioned above, a mixed promoter (also termed “broad with peak”; [17]), an additional promoter type, was revealed. This promoter type exhibits a dispersed initiation pattern with a single strong transcription start site [6, 35] (Fig. 1A). Several studies classified mammalian promoters using alternative criteria [27, 29, 33]. The Ren Lab classified active promoters based on genome-wide ChIP experiments for TFIID and Pol II, as well as H3Ac and H3K4me, regardless of focused or dispersed initiation patterns [33]. Bajic et. al. [29] define four promoter types, based on distribution of dinucleotides over the promoter regions, CpG Islands and TATA boxes. Moreover, Carninci et. al. [27] classified promoters into four groups based on CAGE analysis: single peak, broad shape peak, bimodal/multimodal peak and broad with dominant peak. These studies also challenge the “focused vs. dispersed” classification, as some mouse and human promoters contain both CpG Islands and TATA boxes. A recent comprehensive review [11], which compared genome-wide studies in human and Drosophila, presented another sub-classification of three major types of promoters termed Type I, Type II and Type III. Type I promoters contain TATA boxes and focused TSSs, lack CpG islands and are associated with tissue-specific expression in adult tissues. Type II promoters contain CpG islands and dispersed TSSs. In mammals, type II promoters lack TATA boxes, and in Drosophila they contain DRE, Ohler 1 or Ohler 6 motifs. Genes belonging to this group are associated with broad expression throughout the
organism's life. Type III promoters are associated with developmentally regulated
genes, which in *Drosophila* contain combinations of Initiator and DPE motifs. In
mammals, type III promoters contain large CpG islands.
Taken together, the transcriptional initiation landscape is more complex than the
simple classification of two types of promoters.

### 1.3. Bidirectional and divergent transcription

Another manifestation of the complexity of transcription initiation is the phenomenon
of bidirectional transcription. Bidirectional transcription, which presents two closely
spaced transcription initiation events (within less than 1kb) of head-to-head Pol II
transcripts in both sense and anti-sense orientations, was originally defined for
adjacent head-to-head oriented pairs of protein-coding genes [36]. The relatively
short region that contains the opposite-oriented initiations and separates between
these genes, is often called a “bidirectional promoter” [37]. Experimental and
computational studies have characterized many features of bidirectional promoters. In
general, it is shown that 10%-22% of the genes in mammals are organized in this
manner [38]. Moreover, the bidirectionality was shown to be controlled in a cell-type
specific manner, and these pairs of genes are coordinately regulated ([38] and refs
therein). Hence, bidirectional promoters might have evolved to facilitate the regulation
of transcription of different genes at the same time, and might consist of two separate,
yet dependent, core promoters. Additionally, a computational analysis supports an
evolutionary role for bidirectional promoters in the emergence of novel species-
specific transcripts [39]. Bioinformatics analysis of the distribution of common core
promoter elements (BRE', TATA box, Inr and DPE) and CpG islands at bidirectional
versus unidirectional promoters, demonstrated that while the BRE' is enriched at
bidirectional promoters, the Inr and DPE elements are similarly detected at both promoter types [40]. The TATA box is rare in general, but is enriched in bidirectional promoters of histone genes. Moreover, it was shown that the CpG islands and Sp1 binding sites are common features of most of the bidirectional promoters, compared to unidirectional promoters [41]. Other studies focused on overrepresented binding-sites of different transcription factors, and in some cases - on their influence on the expression of two opposite genes regulated by a bidirectional promoter [38, 42].

Interestingly, another manifestation of bidirectional transcription involving non-coding RNAs (ncRNAs) was recently characterized. Multiple classes of ncRNAs were identified in different organisms (reviewed in [43]). One of these classes is promoter-associated ncRNAs. During the years, classes of promoter-associated non-coding transcripts were discovered in bacteria, yeast, Drosophila, mouse, human and plants ([43-45] and refs therein). Four studies, published back-to-back in 2008, described new classes of promoter-associated ncRNAs in humans and mice [22, 46-49]. These ncRNAs were generally divided into two classes, termed TSS-associated RNAs (TSSa-RNAs) [48] and promoter upstream transcripts (PROMPTs) [47] or upstream antisense RNAs (uaRNAs) [50], which share many features. They are short, present at low abundance and are associated with CpG islands and active-p promoters-related histone marks (H3K4me3, H3ac), but not with elongation-related histone marks (H3K36me3, H3K79me3).

Non-coding antisense RNAs derived from bidirectional promoters have very short half-lives and are barely detectable. Two recent studies have shown that an asymmetric distribution of polyadenylation signals and U1 snRNP-binding sites surrounding TSSs control transcript stability [50-52]. Notably, bidirectional initiation is also a feature of enhancer RNAs (eRNA; see section 7) [53, 54].
The Lis lab has demonstrated that nearly 80% of active genes have bidirectional promoters, suggesting that bidirectional initiation is a general feature of mammalian genomes [22, 55]. Hence, these divergent ncRNAs may be regarded as markers for active promoters of protein-coding genes [22, 46-48, 56]. Duttke et al. have recently analyzed transcription from human promoters in HeLa cells and have classified promoters into three types: unidirectional promoters, divergent promoters (containing an annotated gene in the forward direction and no annotated gene in the reverse direction) and bidirectional promoters (containing annotated genes in both directions) [57]. Surprisingly, they discovered that about half of human active promoters are intrinsically unidirectional. Moreover, the divergent transcripts result from their own reverse-oriented core promoters. The authors suggest that divergent transcription is not an inherent property of the transcription process, but a consequence of the presence of both forward and reverse-directed promoters. This suggestion is in line with the two occupancy peaks observed for each TBP and Pol II by the Lis lab [55]. The Lis lab observed tight spacing (estimated 110 bp) between the forward and reverse-directed promoters [55], whereas the Ohler & Kadonaga labs, observed variable, however larger, spacing between the two [57]. It remains to be determined whether the difference between these findings results from the differences between the different cell lines used or from the analysis methodology.

Despite the impressive discoveries related to bidirectional transcription in the last few years (which highlight the complexity of gene expression), the functional role of short non-coding antisense RNAs still remains elusive. From this point onwards, we only refer to the comprehensively studied focused and dispersed core promoter types.
2. Core promoter elements: the combinatorial code of precise transcription initiation

The Pol II core promoter is composed of short DNA sequences that are referred to as core promoter elements or motifs. The majority of core promoter motifs serve as binding sites for components of the basal transcription machinery, in particular TFIID, which is composed of TATA box-binding protein (TBP) and TBP-associated factors (TAFs), and TFIIB [4, 58, 59].

The basal transcription machinery recruits Pol II to the core promoter that directs the initiation of transcription [4, 6, 10, 60-62]. Nevertheless, there are no universal core promoter elements, and diverse core promoter compositions have been reported [6, 63]. In this section, we will briefly discuss the majority of core promoter elements (schematically depicted in Fig. 1C and D), which have been analyzed in *Drosophila* and mammals, with particular emphasis on their variety and the relations between them.

2.1. The precisely positioned core promoter elements are common in the focused promoters

Early studies from the Chambon lab described the existence of a putative element at the TSS [64]. The function of the initiator (Inr) as a transcriptional element that encompasses the +1 TSS was articulated by Smale and Baltimore [65]. The Inr is probably the most prevalent core promoter motif in focused core promoters [66-68]. It is mainly bound by the TAF1 and TAF2 subunits of TFIID [69-72]. The mammalian Inr consensus sequence is *YYA*$_{+1}$NWYY (IUPAC nomenclature) [73], and the *Drosophila* consensus is TCA$_{+1}$KTY [71, 74]. Inr-like sequences were also identified in *Saccharomyces cerevisiae* [75]. Computational analyses of promoters argue that the
Inr consensus is only YR (-1, +1 positions) in humans [11, 27, 76] or TCA+1GTY for Drosophila [66, 68]. The A nucleotide (or R in the YR consensus) is generally designated as the +1 position, even when transcription does not initiate at this specific nucleotide. This critical convention is instrumental, because functional downstream elements are completely dependent on the presence of an Inr and the precise spacing from it [6, 10, 13].

Notably, a strict version of the mammalian initiator (sINR), which is present in 1.5% of human genes and enriched in TATA-less promoters of specific functional categories, was defined as CCA+1TYTT, with conserved sequences flanking the motif [77]. The sINR motif functions in cooperation with Sp1 and can replace the conventional Inr, but not vice versa. Similarly to the canonical Inr element, sINR is bound by TAF1 and its function depends on it [77]. The YY1 transcription factor binds sINR, but this binding is dispensable for sINR function [77].

In addition to these versions of the Inr, a few core elements that encompass the transcription start site were identified. The polypyrimidine initiator motif (TCT), which was originally identified in mouse, is conserved from Drosophila to humans [14, 78-80]. The TCT has a consensus sequence of YYC+1TTTY in Drosophila and YC+1TYTYY in humans, in which C is the +1 TSS. Although the Inr consensus resembles the TCT consensus, the TCT motif cannot substitute for an Inr to initiate transcription [14]. The TCT overlaps with a motif that was previously identified in humans, termed 5’-terminal oligopyrimidine tract (5’-TOP) (reviewed in [81]), which is functionally distinct from it [14]. Both the TCT and the 5’-TOP elements are enriched and are functional in the transcription of ribosomal protein genes and proteins involved in the regulation of translation [14, 78].
Two additional core promoter motifs that are located around TSSs were originally identified in the hepatitis B virus X gene promoter, which contains two TSSs. The X gene core promoter element 1 (XCPE1) drives Pol II transcription from the first TSS of the X gene promoter as well as from other human promoters, when accompanied by co-activator sites. XCPE1 is found in ~1% of the human genes (particularly TATA-less genes) and its consensus sequence DSGYGGRAS_{+1}M spans positions -8 to +2 relative to the TSS [82]. Unlike XCPE1, The X gene core promoter element 2 (XCPE2) is sufficient to drive Pol II transcription by itself. The XCPE2 directs transcription from the second TSS of the X gene mRNA, but it also drives transcription from additional human promoters, in a TAF-free manner. Its consensus sequence VCYCRTTRCM_{+1}Y spans positions -9 to +2 relative to the TSS [83].

There are core promoter elements that are located upstream of the TSS. The TATA box motif is the first core promoter motif to be identified [84]. Although the TATA box was previously considered to be a universal element, it is presently estimated that only 8%-30% of metazoan core promoters [27, 33, 60, 68, 85] and 20%-46% yeast promoters [62, 86, 87] are TATA-dependent. The TATA box motif is also present in plants [88, 89], however the majority of Arabidopsis promoters are TATA-less [90]. The TATA box is bound by the TBP subunit of TFIID ([5, 6, 63] and refs therein). Both the TATA box element and the TBP are conserved from archaean to humans [10, 91]. The consensus sequence of the TATA box is TATAWAAR, where the 5' T is usually located at -30 or -31 relative to the TSS in metazoans (or at -120 to -40 in yeast). A wide range of sequences can functionally replace the yeast TATA box for in vivo transcriptional activity [92]. Notably, transcription initiation in yeast, which occurs at variable and greater distances downstream of core promoter elements such as the TATA box, has been suggested to occur via Pol II scanning, where a PIC
assembles at the TATA box and Pol II translocates downstream, searching the DNA for suitable start sites (reviewed in [93]). Interestingly, the TATA box has recently been shown to contribute to high rates of transcription re-initiation of human microRNAs, resulting in reduced microRNA levels [94]. The authors suggested that the lower synthesis rates directed by the TATA box result from Pol II crowding at pause sites, thus increasing the chance for premature termination.

The TFIIB recognition elements (BRE), which are bound by the TFIIB basal transcription factor, are located immediately upstream or downstream of the TATA box, respectively [95-97]. TFIIB contacts these two elements by two independent DNA-recognition motifs within its core domain [96]. The consensus of the upstream BRE (BRE\textsuperscript{u}) is SSRCGCC [97], and the consensus of the downstream BRE (BRE\textsuperscript{d}) is RTDKKKK [95]. The TFIIB and the BRE elements are conserved from archae to humans [6, 96]. Both BRE\textsuperscript{u} and BRE\textsuperscript{d} act in conjunction with the TATA box [6, 10]. A bioinformatics analysis using the EPD database showed that 25% of the eukaryotic core promoters contain a potential BRE\textsuperscript{u} [85]. Surprisingly, this study revealed that the BRE\textsuperscript{u} is more prevalent in TATA-less promoters (28.1%) than in TATA-containing promoters (11.8%). Both elements exert positive as well as negative effects on basal transcription and on activated transcription in a manner that is context-dependent [95, 97-99].

In addition to the abovementioned upstream elements there are core promoter elements that are located downstream of the TSS. The downstream core promoter element (DPE), which was discovered as a TFIID recognition site that is downstream of the Inr, is precisely located at +28 to +33 relative to the A\textsubscript{+1} of the Inr, with a functional range set of DSWYVV [100-102]. In addition to this functional range set, the guanine at +24 was shown to contribute to DPE function [102]. The DPE is
prevalent in developmental gene networks [11, 15, 99, 103]. Importantly, a recent study provides in vivo evidence that expression driven by the homeotic Antennapedia P2 promoter during Drosophila embryogenesis is dependent on the DPE [103]. The motif ten element (MTE) was identified as an overrepresented core promoter sequence, which is located immediately upstream of the DPE, encompassing positions +18 to +29 relative to the A+1 of the Inr [68]. As positions +28 to +29 overlap the DPE, the MTE consensus sequence was defined for positions +18 to +27 (CSARCSSAAC) [104]. Although the majority of the MTE-containing promoters contain a DPE, the MTE motif functions independently of the DPE [104, 105]. Both the MTE and DPE serve as recognition sites for TFIID and appear to be in close proximity to TAF6 and TAF9 [101, 105]. Interestingly, TAF1 and TAF2 may also contribute to the recognition of downstream core promoter elements [70]. Using single-nucleotide substitution analysis, the MTE and DPE together were found to consist of three functional sub-regions: positions 18-22, 27-29 and 30-33 downstream to the A+1 of the Inr. The bridge configuration, which includes the first and the third functional sub-regions (bridge I, positions 18-22 with favored nucleotides CSARC; bridge II, positions 30-33 with favored nucleotides WYVY), was shown to be a naturally rare but functional core promoter element [105]. Both the MTE and DPE are conserved from Drosophila to humans [6, 100, 101, 104-108]. The MTE, DPE and Bridge motifs are exclusively dependent on the presence of a functional Inr, and are enriched in TATA-less promoters. However, co-occurrence of putative TATA, Inr and DPE motifs was observed in a small fraction of Drosophila genes [15, 85].

An additional downstream element was identified and characterized in the human adult β-globin promoter. This element, termed downstream core element (DCE), was detected by scanning mutagenesis of the +10 to +45 in the promoter region. The DCE
is composed of three sub-elements, located at positions +6 to +11 (necessary motif CTTC), +16 to +21 (necessary motif CTGT), and +30 to +34 (necessary motif AGC) relative to the TSS. The DCE is distinct from the MTE, DPE and Bridge downstream elements, as the DCE is recognized and bound by TAF1 [109] and not by TAF6 or TAF9 [101, 105]. Moreover, unlike the DPE, the DCE is frequently found in TATA box-containing promoters [109, 110]. Table 1 summarizes the precisely positioned core promoter elements found in focused promoters and the proteins that bind them.

2.2. Core promoter elements with weak positional biases in dispersed promoters

Even though the vast majority of core promoter elements are precisely located in focused promoters, there are still a few variably located motifs that were also identified in dispersed promoters. These variably located elements, like some of the precisely located elements discussed above, are associated with specific gene groups.

As mentioned, there are sequence motifs such as the DNA-replicated-related element (DRE) and Ohler 1, 6 and 7 motifs, which were detected by a computational analysis as commonly expressed in dispersed promoters of *Drosophila* genes with maternally inherited transcripts [28]. The consensus sequences of the DRE, Ohler 1, 6 and 7 motifs are WATCGATW, YGGTCACACTR, KTYRGTATWTTT and KNNCAKCNCTRNY, respectively [68]. The DRE is a target of the DNA replication-related-element binding factor (DREF). DREF, which was discovered in *Drosophila* and was later found to have orthologues in many other species (including humans), is involved in transcriptional regulation of proliferation-related genes [111]. A motif 1 binding protein (M1BP) has recently been identified and the enrichment of Motif 1 and
M1BP was implicated in cytoskeletal organization, mitotic cell cycle and metabolism [112].

2.3. The interplay between core promoter elements

With the notion that there are no universal core promoter elements and that core promoter elements are a very important feature of regulation of gene expression, many studies examined the combinations between core promoter elements such as: Inr, TATA box, BRE\textsuperscript{u}, BRE\textsuperscript{d}, MTE and DPE, and their effects on the transcriptional output. For example, the BRE elements were originally characterized as functional elements with conjugation to TATA box. In this context, both the BRE\textsuperscript{u} and the BRE\textsuperscript{d} either increase or decrease the levels of basal transcription [95, 97, 98, 113]. Notably, the addition of a BRE\textsuperscript{u} element to a core promoter of a Caudal target gene has a differential effect on transcription in a TATA box- or DPE- context [99]. The TATA box and the Inr cooperate, in certain cases, as synergistic elements [114]. An antagonistic behavior was demonstrated between TBP, which activates TATA transcription and inhibits DPE transcription, and NC2 and Mot1, which activate DPE transcription by inhibiting the function of TBP [115].

The functionality of the DPE, MTE and Bridge elements is, by definition, dependent on their precise location relative to the Inr [100, 101, 104, 105]. Synergy was observed between the MTE and DPE, as well as between the MTE and TATA box [104]. Based on these relationships, a synthetic core promoter, termed super core promoter (SCP), containing a TATA box, Inr, MTE and DPE was designed. Remarkably, the SCP is stronger than any of the natural core promoters examined [116].

Collectively, these findings indicate that the levels of gene expression can be modulated by the core promoter composition. Such modulation is directly achieved by
the impact of the combinations of core promoter elements on the architecture of the basal transcription machinery, which provides an additional level of transcriptional regulation. The core promoter may have diversified during evolution so that each element may work with the other, depending on the context and organism. Hence, simple categorization may disregard the complexity of gene expression.

3. Functional and structural insights regarding the role of the core promoter in the assembly of the Pol II transcription machinery

In this section, we describe the assembly of the basal transcription machinery components (primarily based on the analysis of TATA-dependent promoters) and their distinct roles in specific cellular contexts.

3.1. Terminology change: from “general” to “basal” transcription machinery

Classic biochemical studies performed over 30 years ago using the TATA box-containing adenovirus major late promoter identified the general transcription factors (GTFs) as accessory factors for accurate Pol II transcription initiation [117, 118]. The GTFs were named TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH, based on the protein fractions they purified in (reviewed in [4]). These components, together with Pol II, were necessary and sufficient for basal transcription of the adenovirus major late promoter. They assemble into the preinitiation complex (PIC) by protein-protein interactions and by mediating core promoter recognition (Fig. 2B).

In the past, it was generally accepted that the PIC composition of GTFs does not vary between promoters with different core promoter architecture, and the PIC is nucleated by the binding of the TBP subunit of TFIID, which binds the TATA box [119] (reviewed in [4, 31]). Traditionally, this simple model has been considered “general”.
However, due to the diversity in core promoter composition and the realization that the known GTFs are insufficient to transcribe DPE-containing promoters [120], it is suggested that the GTFs do not function in a “general” manner, and different compositions of PIC exist. Indeed, the non-ubiquitous expression pattern of certain TAFs imply that they cannot be PIC components in every cell type [58]. Moreover, many studies have presented the variability in PIC formation, specifically by the molecular flexibility in TFIID composition. Hence, GTFs should be addressed as “basal” rather than “general” transcription factors (also discussed in [58, 121-123]).

3.2. Compatibility between PIC components, related factors and core promoter elements

Undoubtedly, the diverse assembly of the basal transcription factors, as well as the diversity of core promoter elements, is a complex subject, both structurally and functionally. Nevertheless, due to this complexity, the PIC, which is pivotal for core promoter recognition ([58, 121, 124] and refs therein), can assemble at core promoters with varying compositions and regulate Pol II transcription in different cells and organisms. In agreement with that, requirements for a “match” between the PIC and the core promoter have been observed in recent years.

This compatibility has mainly been reflected in studies addressing the flexibility and modularity of TFIID subunits and the entire TFIID complex. Early footprinting assays detected differential TFIID protection patterns with respect to the presence of a TATA box and BRE in mammalian promoters [125, 126], and a DPE in Drosophila [101]. These studies and others [127] have demonstrated the important roles of TAFs in the assembly of the PIC, and hence, in the transcription process. As mentioned earlier, sub-modules of TFIID bind specific core promoter elements, e.g. TBP binds the TATA box, TAF1/TAF2 bind the Inr, TAF1 binds the DCE and TAF6/9 bind the DPE and the
MTE (Fig. 2C) [69-71, 100, 101, 104, 107, 109]. It is noteworthy that TAF4/TAF12 and TAF4b/TAF12 sub-complexes can also bind core promoters [107], and are necessary for transcription of a sub-group of genes, which are mostly associated with TATA box and Inr motifs [128]. Interestingly, TAF1 contains two distinct enzymatic activities: an acetyl-transferase and a kinase activity, which are important for regulating non-overlapping, different gene sets in vivo [129], suggesting that different functional modules of the PIC contribute to transcription of different target genes.

While TBP and TAF1 were initially considered the nucleating subunits of holo-TFIID assembly [130], Wright et. al. [131] discovered that Drosophila TAF4 preferentially nucleates TFIID in TATA-less, DPE-containing promoters. This study also uncovered a stable core-sub-complex, composed of TAF5 and the histone fold domain (HFD)-containing TAF4, TAF6, TAF9 and TAF12. This core sub-complex is associated with the peripheral subunits TAF1, TAF2, TAF11 and TBP. These core TAFs are incorporated into TFIID in two copies, and are organized in five heterodimer pairs with other HFD-containing TAFs (TAF3-TAF10, TAF6-TAF9, TAF4-TAF12, TAF8-TAF10 and TAF11-TAF13) ([124] and refs therein). Recent structural analysis of human TFIID demonstrated that these core TAFs exhibit two-fold symmetry [132]. Interestingly, incorporation of the TAF8-TAF10 pair breaks the symmetry and allows the entry of the single copy TAFs and TBP into the structure, resulting in an asymmetric holo-TFIID that can nucleate the PIC.
Several TBP-free complexes have been characterized [127, 133, 134]. One of them, the TBP-free TAF-containing complex (TFTC; also termed SAGA, STAGA and PCAF), is capable of replacing the canonical TFIID at both TATA-less and TATA-containing promoters \textit{in vitro} [127]. The assembly of TAF-less TBP-containing complexes (such as TBP-TFIIA-containing complexes) at specific core promoters, which was somewhat surprising, has also been observed [135-137]. A TAF-free TBP-containing PIC is important for transcription from HIV-1 LTR promoter [136]. Interestingly, a distinctive TBP-TAF complex, lacking TAF1, TAF4 and TAF10, is involved in transcription of the U2 snRNA gene [138].

These findings add to a growing body of evidence implying that distinct core promoters would be differentially recognized by PICs that contain TBP or are devoid of it. Notably, TBP activates TATA-dependent transcription and represses DPE-dependent transcription, whereas Mot1 and NC2 block TBP function and thus repress TATA-dependent transcription and activate DPE-dependent transcription [115, 139]. Interestingly, Deng et al. [140] demonstrated that NC2 acts positively at promoters that lack functional BREs, while TFIIA recruitment, which is dependent on the presence of BREs, reduces transcriptional activity. The association of BRE elements with TATA boxes further supports these findings [85, 97]. Interestingly, the architectural DNA-binding protein HMGA1 has been shown to interact with the Mediator and activate transcription of mammalian promoters containing both a TATA box and an Inr [141].

Remarkably, the Nogales lab used electron microscopy to visualize human TFIID with promoter DNA, and discovered that TFIID exists in two structurally distinct conformations (termed canonical and rearranged) [142]. The transition between the two states is modulated by TFIIA, and the presence of TFIIA and promoter DNA
facilitates the formation of the rearranged conformation [142]. Human TFIID is comprised of three main structural lobes (termed lobe A, B and C) [142, 143]. Using the super core promoter DNA [116], lobe C was shown to interact with downstream elements (DPE and MTE), while lobe A interacts with the Inr and TATA box.

Three TBP-related factors (TRF1, TRF2 and TRF3) have been discovered in the animal kingdom based on their homology to the C-terminal core domain of TBP, which is essential for interaction with the TATA box (reviewed in [121-123, 144-146]. Unlike TRF1 and TRF3 (also termed TBP2 and TBPL2), TRF2 (also termed TLP, TLF, TRP and TBPL1), is unable to recognize the TATA box, as the TATA-interacting Phe residues of TBP are not conserved in TRF2 [147-149]. Drosophila TRF2 selectively regulates the TATA-less Histone H1 promoter, whereas TBP regulates the TATA-containing core Histones genes [137, 150]. The Kadonaga lab has recently discovered that TRF2, and not TBP, regulates transcription of ribosomal protein genes that lack TATA box and contain functional TCT motifs [151]. Kedmi et. al. [152] discovered that TRF2 preferentially functions as a core promoter regulator of DPE-containing promoters. These findings and others have highlighted the involvement of TRF2 in the regulation of diverse biological processes driven by distinct core promoter compositions (reviewed in [123]). Taken together, promoter recognition by multiple TAFs, TRFs, TBP-free or TBP-containing complexes, underscore a key regulatory role for core promoters in transcription initiation, and may provide an explanation for evolutionary changes affecting the PIC-promoter interface [153].

3.3. Different basal transcription factors promote distinct biological processes

The diversity in the components of the PIC, especially in TFIID subunits, establishes distinct protein complexes that drive transcription of specific sets of genes (e.g. with
cell type- or tissue-specific functions) (reviewed in [154]). The Wassarman lab has shown that *Drosophila* TAF1 affects multiple developmental events *in vivo* [155], and that *Drosophila* TAF6 is broadly required for cell growth and cell fate specification [156]. Moreover, *Drosophila* TAF4 and TAF6 were shown to be required for transcription of the *snail* and *twist* Dorsal-target genes *in vivo* [157]. Human TAF8 was implicated in differentiation of cultured 3T3-L1 preadipocytes to adipocytes [158]. Interestingly, the *Drosophila* TAF10 homologues TAF10 and TAF10b, are differentially expressed during *Drosophila* embryogenesis [159]. Expression of mouse TAF10 was later shown to be required for early mouse embryogenesis of the inner cell mass, but not the trophoblast [160]. Remarkably, conditional knock out of mouse TAF10 in embryonic and adult liver resulted in the dissociation of TFIID into individual components [161]. Based on these findings, it was suggested that TFIID is not required for the maintenance of ongoing transcription of hepatic genes. Rather, it is involved in mechanism of postnatal silencing of hepatic genes [161]. Additional studies reveal an important role for distinct TFIID complexes in regulating pluripotency of embryonic stem cells [162, 163].

Multiple TAF paralogues have been implicated in different biological processes. A retroposed homologue of human TAF1 (TAF1L) and TAF7L are expressed during male germ-cells differentiation [164, 165]. Similarly to humans, TAF7L in mice is required for spermatogenesis in cooperation with TRF2 [165-167]. TAF7L was recently demonstrated to be an important regulator of white- as well as brown-adipose tissue differentiation [168, 169]. TAF4b was originally identified as a cell-type-specific TAF in a human B lymphocyte cell line [170]. Using knockout mice, TAF4b was shown to be important for ovarian development and spermatogenesis [171-174]. Remarkably, mouse TAF9L was recently shown to regulate neuronal gene
expression in vivo [175]. Interestingly, tissue-specific TAF homologues of *Drosophila* TAF4 (*no hitter*), TAF5 (*cannonball*), TAF6 (*meiosis 1 arrest*), TAF8 (*spermatocyte arrest*) and TAF12 (*ryan express*) collaborate to control a testis-specific transcriptional program [176].

TBP paralogues are involved in distinct biological processes, such as embryonic development, differentiation and morphogenesis (reviewed in [121, 123, 145, 177]). TRF2 regulates a subset of genes that differ from TBP-regulated genes. TRF2 is essential for embryonic development of *C. elegans, Drosophila, zebrafish* and *Xenopus* [121, 123, 145, 177]. It is highly conserved in evolution and is present in all bilaterian organisms [147]. Since bilaterian organisms contain three germ layers (endoderm, mesoderm and ectoderm) and more ancient animals only contain two germ layers (endoderm and ectoderm), it is tempting to speculate that TRF2 may be important for mesoderm formation. This suggestion is further supported by the fact that the DPE motif is prevalent among *Drosophila* genes that are involved in embryonic development [15, 99]. Mouse TRF2, unlike *C. elegans, Drosophila, zebrafish* and *Xenopus* TRF2, is not required for embryonic development but is essential for spermiogenesis [178, 179]. A separate study demonstrated that the cleavage of TFIIAα-β precursor (into the α and β subunits of TFIIA) is necessary for activation spermiogenic TRF2 target genes [180]. *Drosophila trf2* is also required for the response to the steroid hormone ecdysone during *Drosophila* metamorphosis [181]. Hence, TRF2 drives multiple transcriptional programs [123].

Zebrafish TRF3 is important for initiation of hematopoiesis during embryonic development [182, 183], however, both zebrafish and *Xenopus* TRF3 are mainly expressed in oocytes and are essential for embryogenesis [184, 185]. Mouse TRF3, which is exclusively expressed in oocytes, is essential for the differentiation of female
germ cells but not for embryonic development [186].

These fascinating findings emphasize the motivation to investigate the regulation of gene expression at the core promoter level. It is possible that there are core promoter motifs that have not yet been discovered, and they might be bound by other PIC components. Thus, the analysis of novel core promoter elements in multiple organisms is likely to shed light on mechanistic aspects of transcriptional regulation.

4. **Enhancer-promoter connectivity**

Zooming out from the basal transcription resolution uncovers another facet of regulation of gene expression, namely, enhancer-promoter interactions that regulate the activation of specific genes in a precise spatio-temporal manner. Enhancers contain DNA binding sites for sequence-specific transcription factors that in turn, recruit co-activators and co-repressors and determine the overall activity of the enhancers (reviewed in [187-194]). Originally, scientists searched for enhancers as cis-regulatory elements that stimulate transcription levels from the nearest promoter, irrespective of orientation. Enhancer-promoter pairs are commonly engaged by enhancer's looping, which physically brings these regulatory elements into proximity, through recruitment of multiple proteins (activators, co-activators, Mediator, cohesin and the PIC). Studies in recent years, employing advanced global methodologies such as chromatin conformation capture (3C), its derivatives (4C, 5C, Hi-C) and ChIA-PET, have led to the discovery of both intrachromosomal and interchromosomal physical contacts with promoters. While multiple enhancers can interact with multiple promoters, specificity between certain enhancers and promoters has been observed. The mechanisms that determine enhancer–promoter specificity are still poorly understood, but they are thought to include biochemical compatibility, constraints
imposed by the three-dimensional architecture of chromosomes, insulator elements, and effects of local chromatin environment [194].

In the last twenty years, the compatibility of enhancer-promoter interactions has mostly been studied in Drosophila. One of the early studies analyzing the compatibility between enhancer-promoter pairs examined the expression of the neighboring gooseberry (gsb) and gooseberry neuro (gsbn) genes [195]. Swapping experiments revealed that although both enhancers (GsbE and GsbnE) are located between the two TSSs of the two genes (and thus cross-activation could potentially occur), the GsbE could only activate the gsb promoter, while the GsbnE could only activate the gsb promoter. Another study showed compatibility between the decapentaplegic (dpp) promoter and its enhancer, which only activates the dpp gene, but not other genes that are located closer to it [196]. Erythroid-specific long-range interactions have been observed in vivo between the active murine β-globin gene and the locus control region (LCR) [197]. These long-range interactions of the β-globin gene were not observed in non-expressing brain cells. High-throughput imaging of thousands of transparent transgenic zebrafish embryos (which were injected with about two hundred combinations of enhancer-core promoter pairs driving the expression of the GFP reporter gene), demonstrated the specificity of individual enhancer-promoter interactions and underscored the importance of the core promoter sequence in these interactions [198]. Taken together, these results demonstrate distinct compatibilities of enhancers to their cognate promoters and the importance of the core promoters in the regulation of enhancer-promoter interactions.

While a few studies in Drosophila demonstrated the involvement of proximal-promoter elements in enhancer specificity [199, 200], there are multiple examples of enhancer-promoter communications that are affected by specific core promoter
elements. Promoter competition experiments revealed that both the AE1 enhancer from the *Drosophila* Antennapedia gene complex and the IAB5 enhancer from the Bithorax gene complex preferentially activate TATA-containing promoters when challenged with linked TATA-less promoters [201]. Nevertheless, both enhancers were able to activate transcription from a TATA-less promoter in reporters that lacked a linked TATA-containing promoter [201]. Enhancer-promoter specificity was first demonstrated in transgenic *Drosophila* sister lines that contain a DPE- or a TATA-dependent reporter gene at precisely the same genomic position relative to the enhancer [202]. Remarkably, this study identified enhancers that can discriminate between core promoters that are dependent on a TATA or a DPE motif. Furthermore, Caudal, a sequence-specific transcription and a key regulator of the *Drosophila* HOX gene network, activates transcription with a preference for a DPE motif relative to the TATA-box [99]. More recently, Zehavi et. al. [15] analyzed the *Drosophila* dorsal-ventral developmental gene network that is regulated by the sequence-specific transcription factor Dorsal, and discovered that the majority of Dorsal target genes contain DPE sequence motifs. The DPE motif is functional in multiple Dorsal target genes, as mutation of the DPE leads to a loss of transcriptional activity. Moreover, the analysis of hybrid enhancer-promoter constructs of Dorsal targets reveals that the core promoter plays a pivotal role in the transcriptional output [103].

High-throughput analyses of enhancers in diverse biological systems have led to a wealth of information with regards to long-range enhancer-promoter interactions and three-dimensional chromatin landscapes. We highlight several remarkable findings below. First, most of the enhancer-promoter interaction loops of regulated genes are distal, and are not localized at the nearest promoter as originally considered [203-205]. Second, enhancer looping enables cooperative regulation of genes of the same
biological process by organizing them in physical proximity [203, 205]. This may indicate a similar core promoter composition among these gene networks or gene clusters (as previously described for the Hox and dorsal-ventral developmental gene regulatory networks [15, 99]).

A recently developed genome-wide screen termed STARR-seq (self-transcribing active regulatory region sequencing) identified thousands of enhancers that could activate transcription of a synthetic promoter containing four core promoter elements in a single promoter - the TATA-box, Inr, MTE and DPE motifs [206]. Notably, enhancers near ribosomal protein genes were under-represented among the enhancers identified in this study, which could be due to the fact that the majority of ribosomal protein gene promoters are regulated via the TCT core promoter element [14, 194, 206].

Remarkably, both the Furlong lab analyzing enhancer three-dimensional contacts during Drosophila embryogenesis, and the Ren lab analyzing long-range chromatin interactions in human cells, discovered that the majority of enhancer interactions remain unchanged during marked developmental transitions or activation following gene induction, respectively [203, 207]. This “on-hold” enhancer-promoter connections, may be preparing the cell for rapid activation of transcription. The Furlong lab discovered that the pre-existing loops are associated with paused Pol II and proposed a model where through transcription factor–enhancer occupancy, an enhancer loops towards the promoter and polymerase is recruited, but paused in the majority of cases (Pol II pausing is discussed below). They suggest that the subsequent recruitment of transcription factor(s) or additional enhancers at preformed enhancer-promoter interaction hubs could trigger activation by releasing Pol II pausing [207]. Notably, enhancer–promoter interactions analyzed in these studies
involve active promoters, with high enrichment for H3K27ac and H3K4me3, and active enhancers, defined by H3K27ac, Pol II and H3K79me3, indicating similarities in 3D regulatory principles from flies to humans [203, 204, 207].

Strikingly, the Stark lab has recently demonstrated that distinct sets of enhancers activate transcription with core promoter specificity using two types of Drosophila cultured cells [208]. They used the core promoter of a ribosomal protein gene driven by the TCT motif, as a representative of housekeeping promoters, and a synthetic promoter (derived from the even skipped promoter), which contains four core promoter elements in a single promoter - the TATA-box, Inr, MTE and DPE motifs, as a representative of developmental promoters. Thousands of enhancers exhibit a marked specificity to one of the two core promoters - the housekeeping promoter or the developmental promoter. Interestingly, TSSs next to housekeeping enhancers were enriched in Ohler motifs 1, 5, 6 and 7 (consistent with the ubiquitous expression and housekeeping functions of these genes), whereas TSSs next to developmental enhancers were enriched in TATA box, Inr, MTE and DPE motifs (which are associated with cell-type-specific gene expression).

Taken together, these observations strengthen the concept that the core promoter composition is not only a pivotal component in basal transcription and initiation, but also an active regulator of transcription that is instrumental for activating developmental and housekeeping gene regulatory programs via sequence-encoded enhancer-promoter specificity.

5. Transcription initiation, Pol II recycling and steps in between: the crosstalk between the core promoter and other modules in the transcription cycle
Apart from transcription initiation, Pol II-driven transcription cycle contains additional steps: elongation and termination. These steps contain at least eight transition points at which transcription is regulated by multiple dedicated factors, and each can be rate limiting (reviewed in [209, 210]). Moreover, maturation of mRNA precursors occurs co-transcriptionally [211]. Below, we briefly describe these highly regulated steps with a focus on the direct or indirect role of the core promoter.

5.1. Timing and synchrony - Pol II pausing and productive elongation

Early elongation, following proper transcription initiation and preceding productive elongation, contains two sequential steps: promoter-escape and promoter-proximal pausing of Pol II. Pol II pausing is a highly regulated step, which is characterized by accumulation of Pol II, typically at 20-60 nucleotides downstream of the TSS (reviewed in [210, 212, 213]). The transition from initiation to early elongation is regulated by multiple factors and phosphorylation events of the heptad repeats within the C-terminal domain (CTD) of the largest subunit of Pol II. The CTD is mostly unphosphorylated when Pol II is recruited to the promoter. Serine 5 (Ser5) of the CTD is then phosphorylated by TFIIH, which causes destabilization of the interaction between Pol II and other PIC components and thus, permits promoter escape and early elongation. Following Ser5 phosphorylation, association of DRB sensitivity-inducing factor (DSIF) and Negative elongation factor (NELF) complexes with the phosphorylated Pol II leads to pausing at the promoter-proximal region [214]. Next, positive transcription elongation factor b (P-TEFb) complex phosphorylates the Ser2 residue of the Ser5-phosphorylated CTD, and the DSIF and NELF factors. These post-translational modifications result in productive elongation (reviewed in [210, 212, 213]).
Pol II pausing was originally identified in *Drosophila* heat-shock and human *c-myc* genes [215-218]. Although Pol II pausing was originally considered to be restricted to a few specific genes, nowadays, the pausing of Pol II appears to be a common step in the transcription process of multiple genes, and generally prevalent in metazoans [22, 219-223]. Specifically, multiple genome-wide assays and studies *in vitro* and *in vivo*, mostly in *Drosophila*, showed that the Pol II pausing has a role in facilitating metazoan developmental control genes and genes that respond to environmental stimuli ([224] and refs therein). Thus, Pol II pausing contributes to developmental dynamics, along with designated transcription initiation programs [225, 226]. Notably, Pol II pausing under normal growth conditions is very rare in *C. elegans* [227] and this is consistent with the lack of a *C. elegans* NELF [214], which contributes to pausing in other organisms. It was previously argued that Pol II pausing prepares genes for a rapid and synchronous induction. Recent studies, however, suggest that paused Pol II is not absolutely required for rapid gene induction, as genes in which Pol II is not paused, can be induced just as quickly, and to even higher levels than paused genes ([213, 224] and refs therein). Promoters regulated by pausing possess a distinct chromatin architecture that may facilitate the plasticity of gene expression in response to signaling events [213]. Notably, paused Pol II complexes were recently shown to be more stable than originally considered, and thus, pausing may serve as a time-window to integrate regulatory signals [228]. There are two known sequence-specific transcription factors that regulate pausing: the GAGA factor (GAF) [215, 216, 221, 229] and the more recently identified M1BP factor [112].

Pausing allows synchronous gene expression of developmentally regulated genes following their induction during embryogenesis [224, 230-233]. Differences in synchronicity are most likely due to the core promoter composition, as demonstrated
by promoter-swapping experiments [231] and the relationship between Pol II pausing and core promoter sequence during *Drosophila* development [230, 234].

The positive elongation factor P-TEFb controls NFκB target genes driven by TATA-containing promoters, whereas the negative elongation factor DSIF controls weak TATA and TATA-less genes [235]. Interestingly, *Drosophila* TATA-dependent promoters are associated with a low degree of pausing [230, 234], suggesting that the TATA box prevents Pol II pausing and promotes P-TEFb activity, leading to a more productive elongation [235].

Remarkably, the Levine lab has shown that at least one fourth of paused *Drosophila* promoters contain a shared sequence motif, the “pause button” (PB), whose consensus (KCGRWC) [236] is similar to that of the DPE (DSWYVVY) [10]. The PB motif is typically located between +25 and +35 (somewhat overlapping the DPE, although it has a wider distribution with regards to its location relative to the TSS). Over one-fifth of the paused *Drosophila* promoters are enriched for the DPE, MTE and PB core promoter motifs, all of which are located close to the pause site [236]. Notably, 75% of the genes in the dorsal-ventral network were identified as paused genes [236]. Over two thirds of Dorsal target genes contain a DPE motif [15]. These correlations, in addition to the fact that PB and DPE are GC-rich and share the ‘GGWC’ sub-consensus, and that both motifs overlap with the paused Pol II (see above), may indicate that the DPE, as opposed to the TATA box, could contribute to Pol II pausing. The Adelman lab has later found out that both the DPE and PB precisely align with the peak of Pol II pausing [222].

In addition, a current study indicates that whereas proximity of Pol II pausing to the TSSs is correlated with focused initiation, pausing at dispersed promoters is located more distally, and with a wider pattern [224, 237]. Moreover, it seems that in contrast
to dispersed promoters, Pol II pausing at focused promoters is not dependent on nucleosome regulation. When the core promoter elements are not located at optimal position, or do not match the consensus sequence, pausing appears to be weaker and located more downstream (+60 to +80) than its typical location. Thus, initiation modes and core promoter architecture affect the strength and location of pausing [237].

It is well known that enhancers play a major effect on activity and synchrony of gene expression in development. Remarkably, Lagha et al. [231] used a promoter swapping strategy and advanced imaging methods and discovered that promoters of key developmental genes play a pivotal role in pausing, which in turn determines the “time to synchrony” - the time it takes to achieve coordinated gene expression in over 50% of the nuclei in the developing Drosophila embryo. The authors demonstrate that substitutions of paused promoters (e.g. tup), which show rapid and synchronous activity, with non-paused promoters (such as pnr), result in slow and stochastic activation of gene expression. Moreover, elements associated with pausing (e.g. GAGA) influence the timing and synchrony of the gene expression. The synchronous activation is essential for proper mesoderm invagination in the developing Drosophila embryo. They provide evidence for a positive correlation between pausing, synchrony and gene expression levels, which are necessary for morphogenesis. Hence, it is the promoter, and not the enhancer, that determines the levels of paused Pol II and the synchrony of gene activation [231, 232].

To summarize, these studies provide evidence regarding different aspects of regulation of Pol II pausing via the core promoter. However, additional biochemical studies are needed to elucidate the mechanisms underlying pausing.
5.2. Termination, polyadenylation and recycling of Pol II - back to square one

The promoter and terminator modules define the boundaries of the transcribed region of protein-coding genes. Transcription termination includes dephosphorylation of the Pol II CTD, its disassociation from the 3’-end and cleavage of the pre-mRNA. Furthermore, this highly regulated event is coupled with the 3’-end polyadenylation processing [238]. Numerous factors in multi-subunit protein complexes and several RNA elements mediate the termination/polyadenylation processes, including two central complexes: cleavage and polyadenylation specificity factor (CPSF) and cleavage stimulation factor (CstF) [239, 240]. Although several factors are shared, the termination mechanism for metazoan replication-dependent core histone genes, which are not polyadenylated, is different than the termination mechanism of polyadenylated genes (reviewed in [239, 241, 242]).

There are mutual links between transcription initiation and termination/polyadenylation. It should be noted that although many studies were done using yeast, we focus here on metazoan transcriptional termination. The CPSF complex was first immunoprecipitated and co-purified with holo-TFIID from nuclear extracts of human cell-lines almost twenty years ago [243]. The authors showed that CPSF is recruited to the core promoter by TFIID and later dissociates from TFIID and continues to be associated with the elongating Pol II and later with the polyA site. Specifically, the CPSF-160 subunit mainly interacts with TAF5, TAF7 and TAF12, but not with TAF1, TAF10 and TAF15 and minimally, if at all, with TBP. Overexpression of TBP reduced polyadenylation of transcripts initiated from a TATA-containing promoter, while both polyadenylated transcripts and non-polyadenylated transcripts that initiated from a TATA-less promoter were unaffected [59, 243]. Furthermore, the recruitment of CstF by TFIIB to the core promoter through PIC assembly was also
demonstrated ([244] and refs therein). Thus, subunits of the main termination factors CPSF and CstF are brought to the PIC and transferred to Pol II, which eventually leads to transcription termination. Moreover, components of the core histone termination machinery were also found associated with histone promoters ([239] and refs therein). Nevertheless, it was previously observed that the termination/polyadenylation machinery influences PIC assembly and the efficiency of transcription re-initiation through Pol II recycling ([245] and refs therein). These transcription initiation-termination/polyadenylation connections are mediated by two different chromatin and genomic mechanisms: gene looping from 3'-end processing sites to core promoters, which brings both modules into spatial and physical proximity, and compartmentalization of genes into “gene factories” [3, 239, 246]. It is noteworthy that these connections and couplings are conserved throughout eukaryotes. In this regard, it is possible that the PIC assemblies and 3'-associated machineries of the core histone genes are particularly specialized, as compared to other protein-encoding genes [137, 239].

In a recent paper, Oktaba et al. [247] demonstrated that the promoters are involved in the regulation of alternative cleavage and polyadenylation. The nuclear RNA-binding protein embryonic lethal abnormal visual system (ELAV) is known to inhibit the canonical polyadenylation processing at the 3' UTRs of genes, which causes to Pol II read-through and 3' UTR extension, during the development of the nervous system in Drosophila and vertebrates. The authors provide evidence that ELAV-mediated 3' UTR extension is dependent on the promoter and Pol II pausing in the developing Drosophila nervous system [247]. Using double-labeling assays and swapping promoters experiments, they show that only reporter constructs that were driven by promoters of known extended genes in vivo, produced extended transcripts
in transgenic *Drosophila* embryos. Ectopic expression of ELAV in non-neural tissues resulted in the induction of 3’ UTR extension. Moreover, sequence analysis of 252 neural-specific transcripts with 3’ UTR extensions revealed the enrichment of the GAGA motif and Pol II pausing. Indeed, reduced 3’ UTR extension levels were observed in GAGA-binding protein Tritorax-like (Trl)-mutant *Drosophila* embryos. ChIP-seq analysis revealed the enrichment of ELAV in promoter regions of extended genes, as well as in 3’ UTRs and introns. Thus, ELAV is selectively recruited to the 3’ UTRs of extended genes through paused Pol II promoters, perhaps via looping between the promoters and the termination regions. Taken together, the above studies strengthen the link between transcription initiation and termination and the pivotal role of the promoter in this linkage.

6. **Is the dogma really composed of sequential steps? – the transcription-translation linkage**

Traditionally, eukaryotic translation has been defined as a separate process that is independent from transcription. However, the translation machinery depends on mRNA-maturation processing, such as the m7G cap structure at the 5’ UTR and its associated protein complexes [248]. These complexes recruit the small ribosomal subunit that in turn reaches the first codon, AUG, via a 5’ UTR scanning mechanism (reviewed in [249]). A common element for translation initiation is the Kozak element (RCCAUGG), which contains the AUG [250, 251]. In addition to this well-defined translational initiator, a distinguished element, Translation Initiator of Short 5’ UTR (TISU), was recently identified. Remarkably, this element is important for transcription and initiation of translation of a specific set of genes [252]. The TISU is found in 4.5% of the mammalian protein-coding genes, with consensus sequence of ‘SAASATGCGCGC’ with rigid core-sequence of 'ATG' located at +5 to +30, and
particularly positioned around the +10 relative to the TSS [60, 252, 253]. This core promoter element is enriched in TATA-less promoters of genes mostly involved in cellular functions such as protein metabolism and RNA processing. As a transcriptional element, it was shown to be necessary for transcription and its function was mediated, at least in part, by YY1 [250, 252]. As a translational element, it was defined as an optimized translation initiator for protein-coding genes possessing a very short 5' UTR (median of 12nt) that mediates translation in cap-dependent but ribosomal-scanning independent manner, as opposed to the Kozak sequence [250, 253]. The 5'-TOP, a mammalian pyrimidine tract regulatory element, was previously characterized as a transcriptional and translational element [78, 79, 254, 255]. It was identified as a core promoter motif used as a transcriptional "initiator" in many protein-biogenesis genes, and its translational activity is critical under stress conditions. The translational control element (TCE) [256], another transcription/translation element, was previously shown to regulate translation in Drosophila testes [257]. Katzenberger et. al. [258] recently showed that the overlapping transcriptional motifs, testis element 1 (TE1) and testis element 2 (TE2), which are overrepresented in testis-specific core promoters, are together identical (TE1/2 motif) to the original TCE. Thus, this element is a transcriptional element, too. The TCE is identified as a transcriptional element in 45% of Drosophila testis-specific genes that are driven by focused promoters. Its consensus sequence is “CTCAAAATTT”, with enrichment in the -5 to +25 region, but without precise location relative to the TSS [258].

Hence, these three core promoter motifs play pivotal roles in both transcription and translation of distinct sets of genes. Moreover, correlations between the TATA box and different features of genes (e.g. gene length) have been observed [259]. This co-regulation of these processes raises questions regarding the interplay between
transcription and translation, such as: Do downstream core promoter elements affect the translation of these genes? Based on the fact that the 5' UTRs of some organisms are short, are these elements evolutionarily conserved? Indeed, a recent study reveals general associations and co-occurrence between translational and transcriptional regulatory trends and features, including core promoter composition [260]. Taken together, the core promoter region is, at least in part, a central intersection for coordinating transcription and translation.

7. Discussion and future perspectives

In this review, we discussed diverse aspects of regulation of gene expression, particularly in metazoans, with an emphasis on the core promoter. We highlighted the complexity of the core promoter architecture. Furthermore, we presented its intricate connections and its pivotal influences on different steps of transcription: initiation, elongation, termination, polyadenylation and finally, translation (Fig. 3). Moreover, we would like to raise a few issues that are directly related to the core promoter but were not mentioned above.

First, in addition to the diversity of core promoter elements and the relationships between them, nucleotide polymorphism in the core promoter affects its activity including its binding by the PIC components. Multiple lines of evidence point towards polymorphisms in many human promoters, particularly in the TATA box sequence. These TATA box substitutions can affect TBP binding and core promoter activity, and are associated with human diseases ([261], reviewed in [262]). It is expected that like TATA box polymorphism, polymorphisms in other elements exist, and may be clinically relevant.

Second, the enhancer-promoter interactome seems to be a much more complex landscape than previously considered. In agreement with that, promoter-promoter
interactions have recently been found [263]. These interactions behave as enhancer-promoter interactions, where one promoter is able to act as an enhancer of another. Hence, hypothetical, more complicated hierarchies of direct and indirect interactions between enhancers and promoters could be achieved (e.g., generating an enhancer-promoter-promoter hub).

Moreover, an additional regulatory aspect that is associated with enhancers is the discovery of enhancer-derived RNAs (eRNAs). This class of ncRNAs was only discovered a few years ago in humans [264]. eRNAs are short-lived, 5’-capped transcripts produced from enhancer regions. Their expression is correlated with histone marks of active enhancers (H3K4me1 and H3K27ac), and they are enriched for transcription factors, co-activators (such as p300/CBP), basal transcription factors and Ser5-phosphorylated Pol II. eRNAs are preferentially found in enhancers that contact their target promoters through enhancer-looping, and it is suggested that these transcripts play a role in generating or maintaining enhancer-promoter-loops and in facilitating the recruitment of sequence-specific transcription factors, chromatin remodeling or chromatin modifying complexes to the targeted promoters [53]. Additionally, eRNAs are associated with several signaling-pathways ([53, 54] and refs therein). Although eRNAs are extensively investigated, also by high-scale methodologies [265], little is known about their core promoter compositions and their TSS architectures [55]. Hence, one of the future goals should be an in-depth investigation of the core promoter architectures of eRNAs and their transcriptional machineries.

Actually, in agreement with the current knowledge that many active mammalian promoters are bidirectional [22, 57], a study published several months ago revealed shared architectures of bidirectional initiations at promoters and active enhancers.
On one hand, similar trends and profiles of transcription factor binding, nucleosome positioning, histone marks and similar frequencies of sequence motifs such as the TATA box, BREs and Inr (YR only) were present in both promoters and transcribed enhancers. On the other, these modules differ in the stability of the transcripts that they synthesize in each direction: promoters give rise to stable transcripts in the sense direction, whereas promoter upstream antisense RNA and enhancer RNAs are rapidly degraded [55]. This unifying architecture of TSSs [266] along with recent findings (e.g. promoter-promoter interactions) challenge the traditional classification of promoters and enhancers (see also [267]). It is noteworthy that Core et. al. [55] indicated that although there are distinct pause modes, which include proximal focused pausing and distal dispersed pausing (see also [237]), the length between the bidirectional TSS pairs and the peaks of TFIIB are not affected. This high-resolution analysis of nascent RNAs might also imply that the high frequency of dispersed mammalian core promoters observed previously, represents multiple independent initiation sites acting as enhancers for neighboring promoters [55]. Thus, the phenomena of dispersed mammalian promoters might be less abundant than originally perceived. Taken together, the growing body of evidence indicates that the core promoter lies at the heart of gene expression.

Acknowledgments

We thank Ron Even for graphic design assistance. We thank Jim Kadonaga, Uwe Ohler, Sascha Duttke, Anna Sloutskin, Hila Shir-Shapira and Racheli Harshish for critical reading of the manuscript. Core promoter-related research in the Juven-Gershon lab is supported by grants from the Israel Science Foundation (no. 798/10), the European Union Seventh Framework Programme (Marie Curie International
Reintegration Grant; no. 256491), the United States-Israel Binational Science Foundation (no. 2009428; joint with James T. Kadonaga) and the German-Israeli Foundation for Scientific Research and Development (no. I-1220-363.13/2012; joint with Eileen E.M. Furlong).
References


Figure legends

**Fig. 1.** General features of the core promoter region. A. The three main core promoter types based on the distribution of TSSs, including focused, dispersed and mixed promoters. Small arrows represent weak TSSs, whereas a large arrow represents a single strong TSS. B. Chromatin features of active core promoters include distinct post-translational modifications and nucleosome depletion. Associated histones marks are depicted: H3K4me2/me3 (orange), H3K4ac (gray), H3K27ac (light blue). A DHS/NDR pattern ranging from nucleosome-free (light) to nucleosome-occupied regions (dark) is illustrated below. C. Schematic illustration of the most common core promoter elements found in focused promoters. The diagram is roughly to scale. D. Schematic illustration of the known factors and sequence motifs that are associated with dispersed promoters.

**Fig. 2.** The core promoter can be studied from different angles in multiple resolutions. A. Zooming in on global genomic interactions in the nucleus, one can study long-range interactions, such as those between enhancers and promoters, by analyzing chromatin looping, cohesion function, interactions of transcription factors (TFs) with co-activators and cis-regulatory modules and interactions of the preinitiation complex (PIC) components with their target promoters. B. Zooming in on the basal transcription machinery, one can study the assembly and composition of the PIC at different Pol II-promoters and on the 3D structure of different PIC components. C. Zooming in on the DNA-binding PIC components (TFIIB and TFIID), one can focus on the alternative protein components at different Pol II-promoters, on the core promoter composition of specialized transcription programs, and on the interactions of different PIC components with specific core promoter elements.
**Fig. 3.** Schematic model depicting the pivotal role of the core promoter module in diverse molecular events and stages of gene expression. The core promoter is important for (clockwise): basal transcription initiation and PIC-core promoter compatibility reflected by different compositions of basal transcription factors, which comprise diverse PIC architectures (top); enhancer-promoter compatibility (which is schematically represented by the preferential activation of DPE-dependent promoters by Caudal); promoter-proximal Pol II pausing (where different core promoter elements support Pol II pausing vs. elongating states); termination/polyadenylation and Pol II recycling (where subunits of the main termination factors CPSF and CstF are recruited to the PIC at the core promoter and continue to be associated with the elongating Pol II and later with the polyA site, which eventually leads to transcription termination. Furthermore, 3’ end formation has been shown to stimulate transcription initiation); and translation (via core promoter elements that play a role in both transcription and translation). Please see the main text for detailed explanations.
Figure 1
Figure 2
Figure 3
Graphical abstract
<table>
<thead>
<tr>
<th>Name</th>
<th>Consensus</th>
<th>Positions (relative to the +1)</th>
<th>Bound by</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRE&lt;sup&gt;a&lt;/sup&gt;</td>
<td>SSRCGCC</td>
<td>Immediately upstream of the TATA box</td>
<td>TFIIB</td>
</tr>
<tr>
<td>TATA box</td>
<td>TATAWAAR</td>
<td>-30/-31 to -23/-24</td>
<td>TBP</td>
</tr>
<tr>
<td>BRE&lt;sup&gt;d&lt;/sup&gt;</td>
<td>RTDKKKKK</td>
<td>Immediately downstream of the TATA box</td>
<td>TFIIB</td>
</tr>
<tr>
<td>Mammalian Inr /sINR</td>
<td>YYA&lt;sub&gt;1&lt;/sub&gt;NWYY /CCA&lt;sub&gt;1&lt;/sub&gt;TYTT</td>
<td>-2 to +5</td>
<td>TAF1 &amp; TAF2 /TAF1</td>
</tr>
<tr>
<td>Drosophila Inr</td>
<td>TCA&lt;sub&gt;1&lt;/sub&gt;KTY</td>
<td>-2 to +4</td>
<td>TAF1 &amp; TAF2</td>
</tr>
<tr>
<td>Human TCT</td>
<td>YC&lt;sub&gt;1&lt;/sub&gt;TYTTYY</td>
<td>-1 to +6</td>
<td>TAF1 &amp; TAF2</td>
</tr>
<tr>
<td>Drosophila TCT</td>
<td>YYC&lt;sub&gt;1&lt;/sub&gt;TTTYY</td>
<td>-2 to +6</td>
<td>TAF1 &amp; TAF2</td>
</tr>
<tr>
<td>XCPE1</td>
<td>DSGYGGRAS&lt;sub&gt;1&lt;/sub&gt;M</td>
<td>-8 to +2</td>
<td>TAF1 &amp; TAF2</td>
</tr>
<tr>
<td>XCPE2</td>
<td>VCYCRTTRCM&lt;sub&gt;1&lt;/sub&gt;Y</td>
<td>-9 to +2</td>
<td>TAF1 &amp; TAF2</td>
</tr>
<tr>
<td>MTE</td>
<td>C5ARCSSAACGS</td>
<td>+18 to +29</td>
<td>TAF6 &amp; TAF9</td>
</tr>
<tr>
<td>DPE</td>
<td>DSWYYVY (functional range set)</td>
<td>+28 to +33</td>
<td>TAF6 &amp; TAF9</td>
</tr>
<tr>
<td>Bridge</td>
<td>Part I: CGANC</td>
<td>Part I: +18 to +22</td>
<td>TAF6 &amp; TAF9 (?)</td>
</tr>
<tr>
<td></td>
<td>Part II: WYGT</td>
<td>Part II: +30 to +33</td>
<td></td>
</tr>
<tr>
<td>DCE</td>
<td>Necessary motifs:</td>
<td>Part I: +6 to +11</td>
<td>TAF1</td>
</tr>
<tr>
<td></td>
<td>Part I: CTTC</td>
<td>Part II: +16 to +21</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Part II: CTGT</td>
<td>Part III: +30 to +34</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Part III: AGC</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. The precisely positioned core promoter elements found in focused promoters and the proteins that bind them.
Highlights - The core promoter: at the heart of gene expression

- There are three major types of transcription initiation patterns in metazoans.
- Diverse PIC compositions are compatible with different core promoter architectures.
- Diverse gene regulatory programs are regulated via enhancer-promoter specificity.
- Core promoter elements affect Pol II pausing, as well as transcription termination.
- Distinct core promoter elements link transcription and translation.