REGULATORY ELEMENTS

Metazoan promoters: emerging characteristics and insights into transcriptional regulation

Boris Lenhard[1,2], Albin Sandelin[3] and Piero Carninci[4]

Abstract | Promoters are crucial for gene regulation. They vary greatly in terms of associated regulatory elements, sequence motifs, the choice of transcription start sites and other features. Several technologies that harness next-generation sequencing have enabled recent advances in identifying promoters and their features, helping researchers who are investigating functional categories of promoters and their modes of regulation. Additional features of promoters that are being characterized include types of histone modifications, nucleosome positioning, RNA polymerase pausing and novel small RNAs. In this Review, we discuss recent findings relating to metazoan promoters and how these findings are leading to a revised picture of what a gene promoter is and how it works.

Transcription start sites (TSSs). Nucleotides in the genome that are the first to be transcribed into a particular RNA.

[1]*Institute of Clinical Sciences, Faculty of Medicine, Imperial College London; and MRC Clinical Sciences Centre, London, UK.*
[2]*Department of Informatics, University of Bergen, Norway.*
[3]*The Bioinformatics Center, Department of Biology and Biotech and Research Innovation Centre, University of Copenhagen, Denmark.*
[4]*OMICS Science Center, RIKEN Yokohama Institute, Yokohama, Japan.*
e-mails: b.lenhard@imperial.ac.uk; albin@binf.ku.dk; carninci@riken.jp
doi:10.1038/nrg3163
Published online 6 March 2012

Gene promoters are the loci overlapping transcription start sites (TSSs), at which the total regulatory input of a gene is integrated into the rate of transcriptional initiation. The immediate role of the promoter is to bind and correctly position the transcription initiation complex, whose main catalytic activity consists of DNA-dependent RNA polymerase. In eukaryotes, RNA polymerase II (RNAPII)-transcribed genes are highly heterogeneous with respect to expression level and context specificity. Therefore, their transcriptional control needs to be highly specialized and dynamic; an important part of this diversity is mediated by different classes of RNAPII promoters, which differ dramatically in their architecture, which in turn determines the promoter function and regulation type[1,2]. We focus here on functional diversity and cross-species equivalence of RNAPII promoters in Metazoa. Early models of promoters and transcription regulation in Metazoa have been inspired by promoter architectures of bacteria and single-cell eukaryotes. However, metazoan promoters are more complex: regulatory elements that control their activity can be spread over larger genomic space, and their number of regulators is higher, reflecting additional and more challenging regulatory tasks faced by multicellular species, such as the development and maintenance of distinct tissues and cell–cell communication.

In eukaryotes, the term 'core promoter' is often used to focus on the DNA region in the immediate vicinity of the TSS, which is assumed to dock the pre-initiation complex (PIC). In the standard view of RNAPII promoter function (FIG. 1a), the core promoter consists of several interchangeable sequence elements around the TSS, which bind core components of the PIC. The core elements (which are modelled as sequence motifs) in Metazoa have been extensively reviewed[1,3–7] and are summarized in FIG. 1b. Despite the similarity of their transcription initiation complexes, different metazoan groups have different key motifs associated with the ubiquitously expressed genes. These differences show that at least some of the motifs are lineage-specific innovations and are not an ancient delineator of promoter types. Alternatively, in some types of promoters, the motifs themselves might not be the major determinants of start site selection. In the classical model, the regulatory input to the core promoter consists of transcription factors binding to sites, either in the promoter region within several hundred base pairs of the TSS (at proximal elements) or further away (at distal elements) (FIG. 1).

There are a number of fundamental questions about metazoan promoters for which satisfactory understanding has not been achieved with this model. In this Review, we first discuss the existence of different promoter classes and their rather surprising correspondence and conservation across different genomes, even across distant metazoans. Although the same functional class might differ substantially with respect to motif and nucleotide composition in distant species, their fundamental functional properties — such as epigenetic

Pre-initiation complex
(PIC). A polypeptide complex consisting of RNA polymerase II and general transcription factors. This forms in the core promoter region around the transcription start site and primes RNA polymerase II for transcription.

B recognition element
(BRE). A core promoter element with consensus sequence SSRCGCC found upstream of TATA box.

modifications and their dynamics, nucleosome configuration and association with long-range regulatory elements — all show clear equivalence. We then turn to other recently discovered properties of promoters for which systematic classification and association with promoter function has not been settled. These include promoter-associated small RNAs and RNAPII pausing, stalling and backtracking at the TSS.

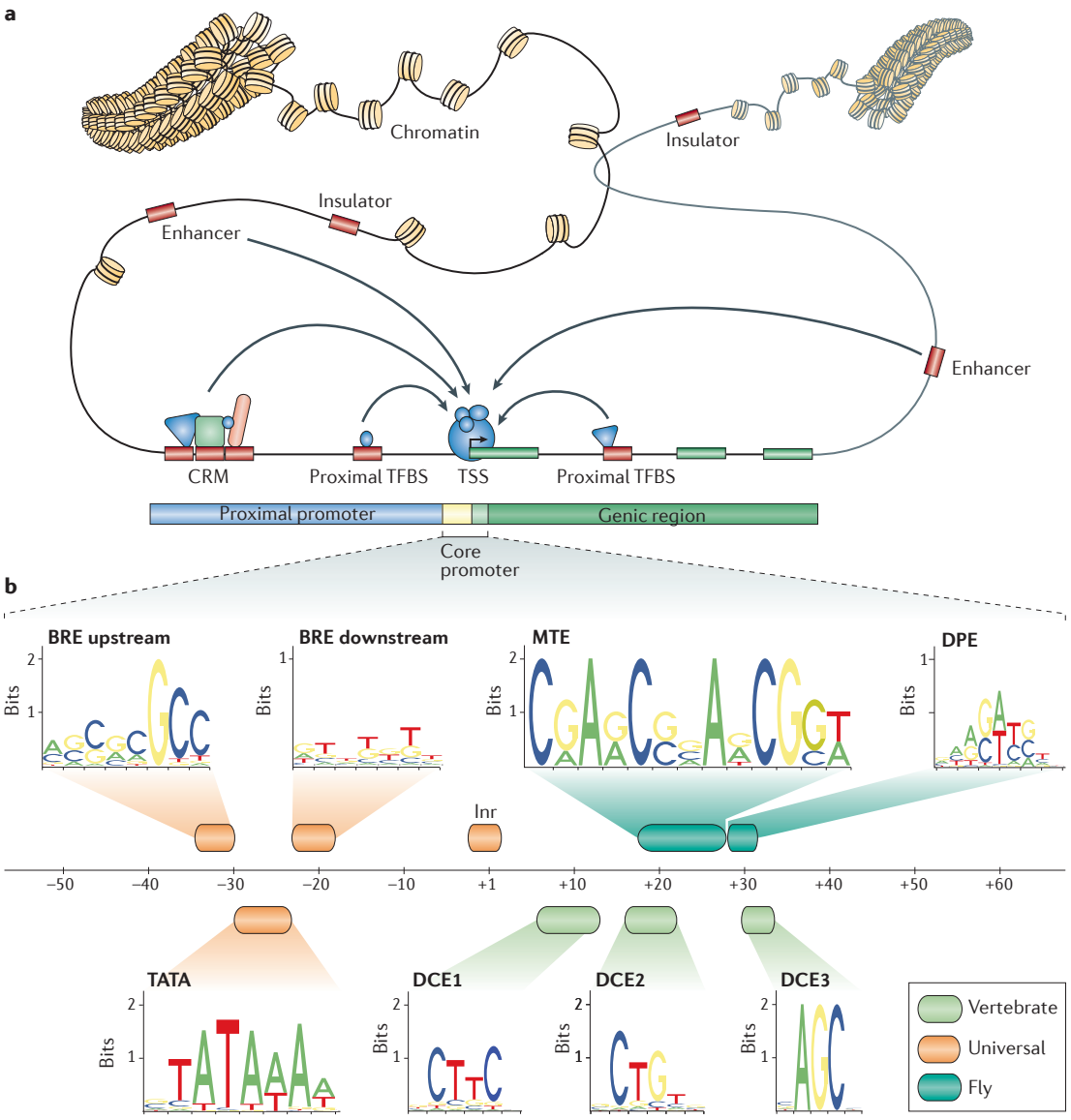


Figure 1 | **Regulation of transcription. a** | A summary of promoter elements and regulatory signals. Chromatin is comprised of DNA wrapped around histones to form nucleosomes. The structure of chromatin can be tightly wrapped or accessible to proteins. Boundaries between these states may be marked by insulators. The region around the transcription start site (TSS) is often divided into a larger proximal promoter upstream of the TSS and a smaller core promoter just around the TSS. The exact boundaries vary between studies. To recruit RNA polymerase II (RNAPII) and to activate transcription of the gene, sequence-specific regulatory proteins (transcription factors) bind to specific sequence patterns (namely, transcription factor binding sites (TFBSs)) that are near to the TSS (proximal elements) or that are far away from it (enhancers). TFBSs can also occur in clusters, forming *cis*-regulatory modules (CRMs). **b** | Sequence patterns in core promoters. The region around the TSS has several over-represented sequence patterns; the TATA box and initiator (Inr) are the most studied and most prevalent. The location of patterns relative to the TSS and their sequence properties are shown as boxes and as associated sequence logos based on the JASPAR database. The Inr pattern is not shown as it varies considerably between studies, ranging from a TCA(G/T)TC(C/T) to a single dinucleotide (pyrimidine (C/T)–purine (A/G)). Importantly, most promoters only have one or a few of these patterns, and some patterns are mostly found in certain species. BRE, B recognition element; DCE, downstream core element; DRE, DNA recognition element; MTE, motif ten element. Figure modified, with permission, from REF. 91 © (2004) Macmillan Publishers Ltd. All rights reserved.

Cap analysis of gene expression
(CAGE). A method for finding transcription start sites.

Chromatin immunoprecipitation
(ChIP). A method for finding DNA–protein interactions that is often combined with sequencing (ChIP–seq) or with microarray analysis (ChIP–chip).

Core promoters: sequence and function

Many of the recent discoveries about transcriptional regulation in Metazoa have been enabled by novel high-throughput-sequencing-based technologies. These methods include sequencing of 5′ ends of mRNAs (cap analysis of gene expression (CAGE) and similar methods) to identify TSSs and chromatin immunoprecipitation (ChIP)-based methods to identify protein–DNA complexes genome-wide (TABLE 1). One of the major discoveries in large-scale detection of promoters was the existence of different classes of core promoters, for which there are common features across the metazoan lineage. The number of main classes has not been settled, but the current evidence points towards three main functional classes (summarized in TABLE 2). In this section, we review the main classes of RNAPII promoters in vertebrates and in *Drosophila melanogaster* — the model invertebrate metazoan in which the promoters and gene regulation have been studied in greatest detail.

Apparent sequence-based dichotomy of promoters in mammals. Initially, the existence of at least two apparent classes of promoters was shown in mammals

Table 1 | **Methods for characterizing promoters and transcription initiation**

Type	Description	Advantages	Disadvantages	Examples
Methods based on sequencing RNA (or cDNA)				
5′-tag-based methods	A range of methods based on capturing capped mRNAs and isolating the first 20–30 nt of corresponding full-length cDNA, which is sequenced and mapped back to the genome. Some methods can identify 3′ ends (paired-end tags) at the same time	These technologies allow single-nucleotide resolution of TSSs (much higher resolution that ChIP-based methods) and thus enable studying the nucleotides of the initiator sites and thus the detailed architecture of the promoter	The methods do not measure promoter activity directly but rather measure the products from transcription, which may be subject to post-transcriptional regulation. Some protocols may require large amounts of RNAs[92], although methods exist to avoid this[65,93]. The methods cannot distinguish capped RNAs and potentially recapped RNA[94]	CAGE[92,93]; nanoCAGE[70]; oligocapping 5′ end SAGE[95]; template switch[65]; 5′–3′ pair end tags (PET)[96]
Small RNA sequencing	Small RNAs (or cDNAs) of a particular size range, often 18–30 nt, are extracted using gel fractionation and are sequenced with next-generation sequencing[97]. The difference between these and 5′-based methods is the size of the RNA and that there is typically no cap selection	Protocols are flexible, as size range, modifications and cell compartments can be selected for. It is possible to select for small RNAs that are bound to specific molecules. As with 5′ tagging, the resolution is very high. A technical advantage is that most sequencing platforms can sequence these RNAs in their full length	As with 5′ tagging, only the result of transcription or RNA processing is measured. As miRNAs and tRNAs or other structure-derived small RNAs[98] dominate the samples, rare small RNAs are harder to characterize and require very deep sequencing	See REFS 81,86,87,99
RNAPII run-on: GRO–seq	This method sequences RNAs that are within elongating RNAPII based on incorporating a label into nascent RNA. The position of elongating RNAPII can be detected	This method can distinguish RNAs that are within RNAPII from other processed small RNAs. High sensitivity	Technically challenging, including isolation of nuclei	See REF. 56
Methods capturing DNA-bound proteins, including RNAPII, transcription factors and histone modifications				
ChIP–seq or ChIP–chip	Proteins bound to DNA are crosslinked by formaldehyde. DNA is then fragmented, and specific protein–DNA complexes can be extracted and isolated using antibodies. DNA fragments are sequenced from their 5′ ends and mapped to the genome (using ChIP–seq) or identified by microarray (using ChIP–chip)	Captures DNA-bound proteins 'in the act' — includes transcription factors, modified histones or proteins that are part of the transcription machinery, such as RNAPII	Heavily reliant on antibody specificity and quality; this makes comparisons between different ChIP experiments hard, even for the same target. The resolution is much lower than for RNA-based methods, as the ChIP fragments are much larger than the bound site. Most analysis also requires computational peak-callling, which has no general standards	A user manual of several technologies (ChIP–seq, DNase hypersensitive site analysis and DNA methylation analysis) was published by the ENCODE Consortium and provides an overview of these methods and applications[100]
Methods locating transcription 'bubbles'				
Permanganate footprinting	Identifies DNA regions that correspond to the 'melted' transcription bubble and can therefore locate the exact genomic sites at which active RNAPII is located	This method detects active RNAPII and not the products of RNA. This method is best at identifying poised RNAPII	At present, it can only be done by targeting specific genes, so throughput is limited	See REFS 101,102

CAGE, cap analysis of gene expression; ChIP–chip, chromatin immunoprecipitation followed by microarray; ChIP–seq, chromatin immunoprecipitation followed by sequencing; GRO–seq, global run-on followed by sequencing; RNAPII, RNA polymerase II; SAGE, serial analysis of gene expression; TSS, transcription start site.

Table 2 | **Promoter types**

Promoter type	Dominant gene function	Common properties	Vertebrate-specific	*Drosophila melanogaster*-specific	Refs
Major promoters					
Type I ('adult')	Tissue-specific expression in adult peripheral tissues	Sharp ('focused') TSS, TATA-box enrichment, disordered nucleosomes	Mostly no CpG islands		8,9,13,17
Type II ('ubiquitous')	Broad expression throughout organismal cycle	Broad ('dispersed') TSS, ordered nucleosome configuration	CpG islands, TATA-depleted	Enrichment of non-positionally fixed motifs (Motif 1 or 6, DRE)	8,9,13,17
Type III ('developmentally regulated')	Differentially regulated genes, often regulators in multicellular development and differentiation	Polycomb repression-regulated genes, broad H3K27me3 marks	Large CpG islands extending into the body of gene	Enriched for DPE	16
Minor promoters					
TCT promoter	Highly expressed genes of translational apparatus	Sharp, pyrimidine-stretch ('TCT') initiator sequence, often full TATA box, ubiquitous-promoter-like nucleosome configuration	CpG island overlapping		23

DPE, downstream promoter element; DRE, DNA recognition element; H3K27me3, histone H3 lysine 27 trimethylation; TSS, transcription start site.

genome-wide[8]. Their main distinguishing feature was GC content and CpG dinucleotide frequency. We shall refer to them as high-CG and low-CG promoters.

The high-CG promoters are characterized by their overlap with a CpG island[6]. The high-CG promoters are mostly associated with multiple TSSs ('broad' promoters) and have been associated with widely expressed or developmentally regulated genes[8,9]. Individual instances of 'broad' promoters were described long before the genome-wide studies (for example, see REF. 10).

Most low-CG promoters exhibit a precise start site at which most transcription initiates from a single nucleotide position (referred to as 'sharp', 'focused' or 'peaked' promoters). Many, but by no means all, have a TATA box at a constrained distance from the TSS. The narrow transcription initiation span is explained by the fact that PIC uses an initiator site at a fixed distance from the TATA box — that is, the TATA box distance in combination with the initiator consensus sequence is the major determinant of TSS selection[11]. TATA-box promoters are associated with tissue-specific transcription, and promoters of many of the genes that undergo lineage-specific family expansions and pseudogenization belong to this class: for example, those of liver-specific genes or olfactory receptors[12].

However, the distinction between these two original classes (the high- and low-CpG promoters) is not razor-sharp and has recently been challenged to an extent by the demonstration that dividing promoters into 'sharp' and 'broad' provides a better functional division of promoter types than a CpG versus non-CpG distinction[13] (discussed in detail below). Some promoters contain both a functional TATA box and a CpG island, and there are indications that such promoters are capable of both TATA-dependent and TATA-independent transcriptional initiation[11].

Classes of promoters in* D. melanogaster*: functional tripartition. In *D. melanogaster*, a number of different promoter types have been suggested based on motif content. An exhaustive analysis of motif composition in *D. melanogaster* and human promoters[14] revealed extensive differences in the type and directionality of motifs found in different promoters and their association with gene function. In parallel, five principal motif-based classes of *D. melanogaster* promoters were proposed[15], which could be further grouped into three general functional classes[16]. For clarity, in this Review, we shall refer to these three classes as types I, II and III. Type I consists of the tissue-specific promoters, which are similar to the low-CpG class in mammals with respect to motif composition, stage of development at which they are expressed and tissue specificity, and they are characterized by a high enrichment for a TATA box at an appropriate distance from an initiator element (Inr element). Type II promoters are associated with 'housekeeping' genes and genes that are regulated at the level of individual cells; they have either a DNA recognition element (DRE) or a combination of novel motifs[15]. Finally, type III promoters have an Inr element only or an Inr element plus a downstream promoter element (DPE). These promoters are preferentially associated with developmentally regulated genes, the expression of which is precisely coordinated across different cells in a tissue or anatomical structure[16].

Promoter class tripartition across Metazoa: unification from CAGE patterns and gene function. Mammalian TATA-enriched, low-CpG promoters are clearly structurally and functionally equivalent to TATA-box-enriched promoters of tissue-specific genes in *D. melanogaster* (type I promoters). Initially, it was less clear whether, in mammals, it was possible to distinguish the promoters of ubiquitously expressed genes and genes that are regulated in development or

CpG island
Genomic sequences that are not depleted of CG dinucleotides, which occurs by 5-methylcytosine deamination. They often overlap or are near to transcription start sites. Most definitions set a minimum length (for example, 200 or 500 bp) and a minimum observed/expected CpG ratio.

TATA box
A T/A-rich sequence that lies upstream of TSSs.

Initiator element
(Inr element). A sequence pattern overlapping the TSSs.

Downstream promoter element
(DPE). This has the consensus sequence RGWCGTG and is common in *Drosophila melanogaster* genes 25–30 bp downstream of the transcription start site.

differentiation, as they both tend to have high-GC, CpG-island-overlapping promoters and a low incidence of TATA boxes. Recently, however, features have been identified that distinguish them[17], and the list of features is still growing. At ubiquitously expressed genes, there is usually only one short CpG island that overlaps with the TSS, and we refer to these as type II promoters. By contrast, developmental genes have several large CpG islands that often extend well into the body of the gene; we refer to promoters of these genes as type III promoters. Mammalian type II and type III promoters also exhibit systematic differences in motif composition and width of the transcription start region, as well as differences in epigenetic features (discussed below).

Remarkably, recent expressed sequence tag (EST) analysis[18] and mapping TSSs using CAGE (TABLE 1) in total RNA from *D. melanogaster* embryos at a series of developmental time points[19] showed that 'sharp' and 'broad' patterns of transcription initiation — a key feature distinguishing vertebrate promoter classes — also exist in *D. melanogaster*. This was surprising for two main reasons. First, early results had indicated that broad promoters were tightly associated with CpG islands, which are not present in the *D. melanogaster* genome, and were considered to be an intrinsic property of such promoters. Second, early results had also indicated that all classes of *D. melanogaster* promoters had well-defined motifs that should constrain TSSs to a small number of initiator positions at a fixed distance from those motifs. It turned out, however, that most 'classical' promoter elements in *D. melanogaster* (such as TATA and DPEs and the recently postulated 'pause button'[20]) are associated with peaked, context-specific (type I) promoters, whereas the broad type II promoters of ubiquitously expressed genes are associated with DREs and a range of weaker, less well-characterized motifs[14,15]. Type III promoters in mammals are, on average, 'sharper' than type II promoters; the difference between the two equivalent promoter classes in *D. melanogaster* is even more pronounced[19].

Interestingly, it also seems now that CpG islands are unlikely to be a requirement for broad promoters in vertebrates[13]. There are sharp promoters that overlap CpG islands, as well as broad promoters that are devoid of CpG enrichment. Instead, it has been shown[13] both in *D. melanogaster* and in humans that sharp and broad promoters have distinct patterns of nucleosome positioning and histone modification that distinguish these promoters much more precisely than the presence of CpG islands (see below). However, it was recently shown that introducing an artificial CpG island into mouse cells leads to the establishment of epigenetic patterns that are characteristic of promoters[1,2] — these patterns are typical of most CpG islands, arguing that in mammals CpG islands are primed to be promoters by default. Because *D. melanogaster* lacks discernible CpG islands, this might mean that other sequence determinants can play the same part in this species. Thus, the quest to determine the key elements that underlie transcription initiation precision is still ongoing.

Expressed sequence tag (EST). An older method that sequences parts of full-length RNAs.

In this Review, we use the results from mammals and *D. melanogaster* to show the functional equivalence of the main promoter classes, because these were the species in which systematic genome-wide TSS data (from CAGE and paired-end analysis of TSSs (PEAT) (TABLE 1)) and extensive epigenomic profiling data have lent support to multiple aspects of this equivalence. However, similar properties have also been observed in other vertebrates (for example, in frogs[21]), and the corresponding classes of promoters can be discerned in a recent study on *Caenorhabditis elegans* promoters[22], even though this species includes unique features, such as *trans*-splicing.

'High-performance' promoters of genes involved in transcription and translational machineries. The promoters of genes for ribosomal proteins and major translation initiation and elongation factors have a combination of distinct features that might warrant their classification into a separate class. The initiator sequence is unlike that of the three main promoter types — it consists of a stretch of pyrimidines and has been termed the TCT motif[23]. The transcribed part of the initiator sequence was subsequently suggested to have a role in coordinated translational response of these mRNAs to amino acid starvation[24]. The TCT promoters are 'sharp', but they differ substantially from type I promoters in other features. In *D. melanogaster*, they typically lack either a TATA box or a DPE[23]. By contrast, their mammalian counterparts usually have a TATA box (which is a common type I feature), but they also overlap CpG islands[25] and show ordered nucleosome positioning, which are features that are shared with most type II and III promoters[13] (discussed below). In each species, this specific combination of promoter determinants apparently enables high-level constitutive expression levels and their coordination in all cell types.

Core promoters: chromatin

In the previous section, we discussed classes of promoters from a sequence-based perspective; however, epigenetic signals — namely, histone and DNA modifications — have been associated with promoter class and functional state. Recent epigenomic data from ENCODE, modENCODE and a number of smaller-scale studies provide additional support for the tripartition of the main functional classes of promoters in vertebrates and *D. melanogaster* (summarized in FIG. 2; details in TABLE 2).

Epigenetic signatures of main promoter classes across Metazoa. Genes that are specifically expressed in peripheral terminally differentiated tissues — such as in liver or skeletal muscle in mammals[8] or cuticle-forming epidermis and endocrine tissues in *D. melanogaster*[16] — have type I promoters with a pattern of histone modification that is distinct from that of most other genes. Histone H3 lysine 4 trimethylation (H3K4me3) is generally only present downstream of the TSS, and there is no RNAPII binding at these promoters when the genes are not active[26,27]. Second, ubiquitously expressed genes have H3K4me3 throughout their type II promoters across all

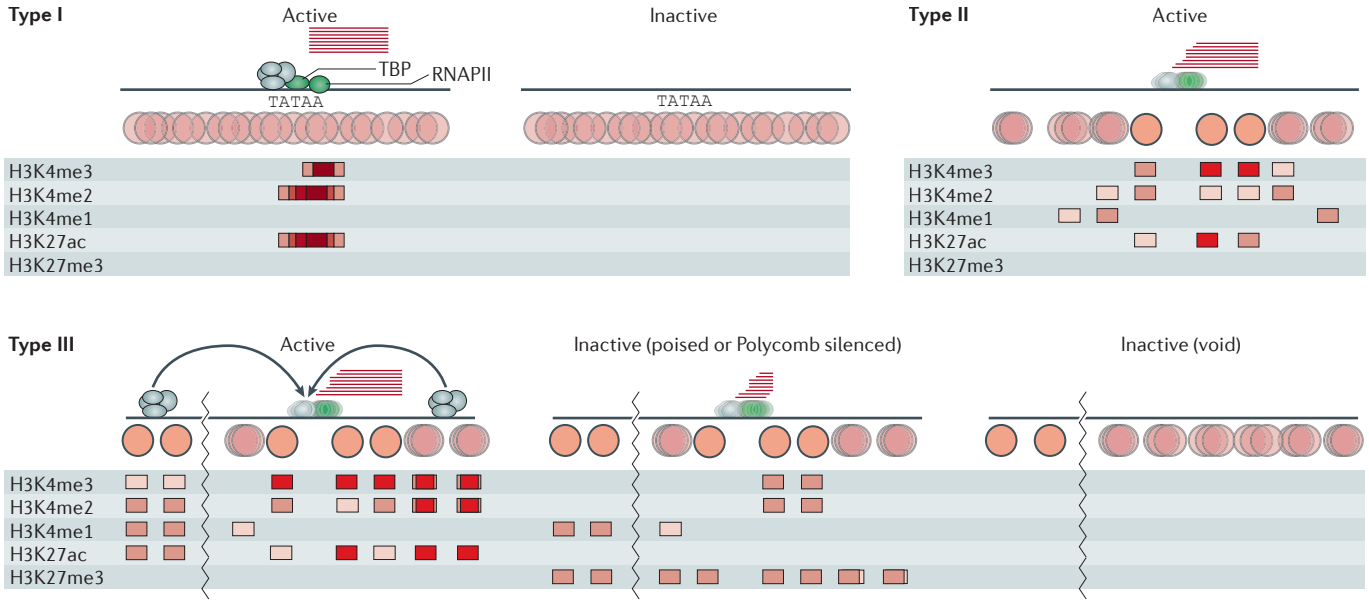


Figure 2 | **Features of the main functional classes of metazoan promoters.** Based on the configuration of promoter signals, transcription start site (TSS) positions, nucleosome positions and their epigenetic marks, most metazoan promoters can be categorized into three general types. This classification represents a broad generalization and should still be considered to be a work in progress. Type I ('adult') promoters are most often used for genes that are specifically expressed in terminally differentiated peripheral tissues of an adult. Type II ('ubiquitous') promoters are active in all cell types. Type III promoters are characteristic of genes with expression that is developmentally regulated and coordinated across multiple cells. TABLE 2 gives more detail about each of the three main types. The DNA is the horizontal black line; the red lines correspond to 5′ ends of individual transcripts and indicate different TSS precision in different promoter types. Sequence-specific transcription factor complexes are in grey, and general transcription factors are in green. The 'fuzziness' represents postulated RNA polymerase II (RNAPII) positioning for TATA-independent initiation on 'broad' promoters. The complexes by no means represent a comprehensive inventory of components at core promoters. Beneath each promoter, nucleosomes are represented by red circles; the 'fuzziness' represents the degree of nucleosome positioning. Histone modifications are shown below the nucleosomes, and the depth of colour represents the prevalence of the modification (with red being the most prevalent). For type III promoters, scenarios with poised RNAPII and/or Polycomb silencing and without RNAPII are shown. H3K27ac, histone H3 lysine 27 acetylation; H3K4me1, histone H3 lysine 4 methylation; TBP, TATA-box-binding protein.

tissues. Across classes of genes in vertebrates, H3K4me3 distribution is almost identical with the span of CpG islands[6]. Ubiquitously expressed genes generally have short CpG islands, and the H3K4me3 mark and CpG island typically only overlap the 5′ end of the gene[17]. Third, developmentally regulated genes (with type III promoters) in vertebrates have a number of features that are associated with repression by Polycomb group proteins (PcG proteins). These features include multiple large CpG islands, wide distribution of bound PcG proteins and both H3K27me3 and H3K4me3. Because of the presence of both of these marks, which are associated with repression and activation, respectively, these are described as bivalent promoters[28]. The large CpG islands often extend into the body of the genes and are closely tracked by H3K4me3, which is thus not restricted to promoter regions in developmental genes. In *D. melanogaster*, broad H3K27me3 and Polycomb repressive complex 2 (PRC2) marks are also present, even though CpG islands are absent, and the existence of bivalent promoters is still unproven. An intermediate 'balanced' state of chromatin[29], which includes a combination of features that are associated with repressed promoters (such as H3K27me3 and Polycomb) and active promoters (such as RNAPII and the Trithorax protein ASH1, but not H3K4me3), could be the functional equivalent of bivalent promoters in *D. melanogaster*.

Several recent studies[26,30,31] have made attempts to classify chromatin based on combinatorial content of multiple epigenetic marks and/or transcription factor binding. The results tell us much about the functional states of promoters across different states and in genes of different types. Although this classification of epigenetic states is cell-type-specific, when it is inspected across many cell types, the three main types of promoters are preferentially associated with different subsets of epigenetic states. For example, in humans, differences have been reported in the relative contributions of enhancer- and promoter-based regulation in different gene classes that clearly track the three promoter types[30]. Developmental genes (type III promoters) are regulated at both the enhancer and the promoter level, have the highest number of enhancers nearby and have diverse promoter states, including poised and Polycomb-repressed states. Tissue-specific genes (with type I promoters) seem to depend predominantly on *cis*-regulatory modules for regulation and show less diversity across promoter states (typically assuming either

Polycomb group proteins
(PcG proteins). These are epigenetic regulators of gene expression that silence target genes by establishing a repressive chromatin state. Because of their role in maintaining states of gene expression, PcG proteins have key roles in cell fate maintenance and transitions during development.

Polycomb repressive complex 2
(PRC2). A regulatory complex that catalyses trimethylation of histone H3 at lysine 27.

Trithorax protein
Proteins that belong to the Trithorax group (TrxG) form large complexes and maintain the stable and heritable expression of certain genes throughout development.

inactive or active state). Finally, the housekeeping genes with type II promoters have few enhancers nearby and are generally characterized by an active promoter configuration. It must be noted that even for the most studied epigenetic marks, only a correlation between active or repressed transcription is known, and knowledge of the interdependence of these marks remains sketchy; causality in terms of knowing what event triggers the next one is only understood for a small subset[32,33]. Further knowledge of the relationships among marks and their deposition, erasure and transmission through cell division would aid in our understanding of the mechanism of promoter action.

Nucleosome occupancy and positioning at promoters. The nucleosome occupancy of a DNA location is measured by determining the proportion of the copies of that sequence in a sample that is nucleosome-bound rather than nucleosome-free, often based on DNA digestion combined with ChIP-based sequencing. Nucleosome positioning refers to the precise preferred position of a nucleosome with respect to the underlying sequence. ChIP-based methods can measure either or both, depending on the resolution and design. Because the number of nucleosomes is large, the sequencing depth constraints generally do not allow analysis of individual core promoters. With the most recent generations of high-throughput sequencing platforms and improved nuclease-based nucleosomal DNA preparation protocols, we expect that this will no longer be a limitation.

Different classes of promoters seem to have different patterns of nucleosome occupancy and precision of positioning. Perhaps counter-intuitively, the 'broad' promoters of housekeeping genes are characterized by a more precise nucleosome positioning than the promoters of tissue-specific genes, which have precise TSSs[13,34]. In both *D. melanogaster* and mammals, broad promoters have a nucleosome-free region that encompasses the TSSs and the immediate upstream region in which multiple TSSs can be used; they often exhibit DNase hypersensitivity, even when the gene is not expressed. Conversely, sharp promoters often have the TSS covered by a nucleosome but have less-ordered nucleosomes flanking this position. In these analyses, the authors did not separate the type II and III promoters as we understand them now.

In addition, recent work[13] has shown that a disordered or ordered configuration of nucleosomes is one of the most striking observable differences between sharp, tissue-specific promoters and broad, ubiquitous promoters. This raises the question of how the nucleosome configuration is established and what role it has in promoter function. The features that influence nucleosome position have been studied in depth, but they vary among organisms[35,36]; further work is required to understand their links to promoter function[37].

Chromatin remodelling. The re-interpretation of the original classification of sharp or broad promoters through the differences in epigenetic and nucleosome configuration properties of different promoter classes sheds new light on studies of the role of chromatin remodellers at promoters. In Toll-like receptor 4 (TLR4)-induced mouse macrophages, the activation of most primary response genes that have CpG island promoters is SWI/SNF-independent, whereas most promoters with SWI/SNF-dependent activation lack CpG islands[38]. This was shown to be linked to the differences in nucleosome organization that distinguish the non-CpG 'sharp' promoters (type II) from the CpG island 'broad' promoters (types II and III, which were not considered separately). Recently, BRG1 — the enzymatic motor component of the SWI/SNF complex — has been shown to help RNAPII overcome nucleosomal barriers during transcription elongation, suggesting that most CpG island promoters might not have such barriers[39].

However, some non-CpG promoters in mice were shown to be SWI/SNF-independent and yet do not have constitutively active chromatin; these promoters are mostly associated with inducible immune system genes[38]. In addition, 12% of primary response CpG promoters were also SWI/SNF-dependent. Nucleosome position and epigenetic properties might be more predictive of SWI/SNF dependence (and vice versa) than the presence of CpG islands. It has been suggested[40] that SWI/SNF dependence or independence is the key difference between primary and secondary response genes, respectively. The reinterpretation of these findings in terms of three promoter classes will probably reconcile some of the ambiguities in this division, and primary-response genes will mainly fall into the type III (developmental) class, whereas secondary-response genes will fall into the type I class (that is, the tissue-specific, terminally differentiated class).

Nucleosome positioning can also be influenced by other DNA-binding proteins. For example, CCCTC-binding factor (CTCF) has been shown to impose positioning on up to ten nucleosomes flanking its binding sites[41]; this way, CTCF may have the ability to influence promoter activity over distances of several kilobases[41].

Three-dimensional interactions and nuclear localization. Different promoters respond differently to long-range regulation. In *D. melanogaster* and vertebrates, developmental genes are more likely to be associated with multiple long-range enhancers and with highly conserved non-coding elements[16,42,43]. ENCODE data have confirmed the correlation across different tissues between the expression level of these genes and the presence of long-range enhancers with active marks[44]. Unlike developmental genes with type III promoters, many tissue-specific genes with sharp, low-CpG (type I) promoters in vertebrates have their key regulatory inputs close to their promoters[45]. The broad dispersal of context-specific regulatory elements around the TSSs of developmentally regulated genes calls into question the common practice of searching for context-specific regulatory signals in all promoters of sets of co-regulated genes. The difference between the type I and III promoters in the typical distance of regulatory inputs from the promoter is visible even when the genes are regulated by the same transcription factor complex[46] and may, in

Nucleosome occupancy
A measure of the degree to which a certain DNA region is bound by a nucleosome.

Nucleosome positioning
The pattern of nucleosome occupancy along DNA.

SWI/SNF
A protein complex that can alter the positions of nucleosomes. It has ATP-dependent chromatin remodelling activity.

CCCTC-binding factor
(CTCF). A transcription factor, one role of which seems to be to define some chromatin boundaries that are associated with differential DNA accessibility.

part, reflect the differences in nucleosome organization between the promoter classes.

The long-range contacts — which can be up to megabase distances — between genes and regulatory elements that target them have been demonstrated experimentally using a range of chromatin conformation capture (3C)-based methods[47]. Also, genes can alter their nuclear localization depending on their activity state. For example, silenced genes are frequently associated with nuclear lamina[48,49] and, following activation, they can relocate to 'transcription factories'[50,51]. Promoters that are specifically regulated during differentiation are therefore likely to move within the nucleus.

New insights into mechanistic complexity

Genome-wide analyses of promoters have also revealed surprising new features, including the dynamics of RNAPII and small RNAs associated with transcription.

RNAPII dynamics: stalling, poising and backtracking. A common finding from RNAPII ChIP experiments has been a clear enrichment of RNAPII occupancy at TSS regions in *D. melanogaster*[27] and mammals[52,53] compared with the gene body. This suggests that RNAPII is frequently stalled near promoters (reviewed in REF. 54). This is either a regulated blockage of transcription until a release or activation signal is received (referred to as 'poising'), or it is an accumulation of RNAPII at the promoter of actively transcribed genes that is due to RNAPII slowing down immediately downstream of the TSS (most often referred to as 'pausing'). This RNAPII enrichment at TSSs was first seen for heat shock genes in several species (for example, see REF. 55), and subsequent genome-wide approaches showed that >55% of non-expressed genes in mouse embryonic stem cells (ESCs) have an accumulation of RNAPII at their promoter. Similarly, analysis of mouse ESCs using a genome-wide run-on assay (such as global run-on followed by sequencing (GRO–seq; TABLE 1) showed that >40% of all genes have an over-representation of RNAPII at the 5′ ends versus the gene bodies[56].

Even though the H3K4me3 mark is generally associated with active promoters, it is also present on promoters with non-elongating RNAPII. This means that H3K4me3 is not necessarily an indicator of active transcription of a gene.

The functional tripartition of promoter types is mirrored by the RNAPII occupancy pattern. In early *D. melanogaster* embryos analysed by ChIP followed by microarray (ChIP–chip)[27], the tissue-specific promoters show no RNAPII binding (they are only activated later in development), and the housekeeping genes show evidence of active transcription by having RNAPII signal along the entire gene body (discussed below). Genes encoding developmental regulators had RNAPII accumulated at their promoters or immediately downstream of them. However, although this distinction is clear in ESCs, in more differentiated cell types, such as in fibroblasts, there seems to be increased paused RNAPII at genes that encode otherwise highly expressed components of the translational machinery[56]. It should be noted that even highly transcribed genes (with successful elongation) have an RNAPII density in promoter regions that is around nine times higher than it is in gene bodies, although this can vary considerably. Therefore, identifying 'poised' or 'stalled' genes depends on the applied thresholds, and it is harder to make this classification for genes with an overall lower transcription level.

The biological function of RNAPII pausing or stalling is unclear. For developmental genes (type III promoters) and heat-shock genes, poising may be a strategy for starting transcription rapidly in response to stimuli. For genes with broad promoters (that is, of either type II or type III), the poising or other kind of stalling might simply reflect open chromatin. Alternatively, it may be that early elongation is the least efficient phase in transcription, and thus the accumulation of RNAPII is simply due to kinetics. In poised genes, elongation might be actively regulated, so RNAPII can be released to achieve bursts of transcription. An additional methodological difficulty — and a mechanistic opportunity — in studying RNAPII dynamics is posed by the fact that RNAPII changes post-translational modifications as it proceeds from initiation to pausing to elongation[57,58], and this can affect antibody recognition specificity in the ChIP experiments. These questions are the subject of active research; to solve them, approaches that can isolate and analyse single cells might be needed.

RNAPII can also move back towards the TSS — so-called 'backtracking'. It has been observed in *D. melanogaster*, both *in vitro*[59] and *in vivo*[60], in which sequence features in the promoters seem to direct stalling and backtracking. Promoters that are enriched for RNAPII by this mechanism are most often associated with developmentally regulated genes (type III promoters) and have an over-representation of DPE patterns[61], and the regions that are just upstream of backtracked RNAPII are more AT-rich than the downstream regions. The lower melting temperature of AT-rich sequences probably favours backtracking[60]. Backtracking has not been as clearly demonstrated in mammals: there, DPEs are not clearly detected, and the regions just downstream of the TSSs are generally not AT-rich, making this backtracking mechanism less likely.

Retrotransposons and repeats recruited as promoters. In the past few years, it has become apparent that transcription can derive from regions that were not previously considered in transcriptome studies. An example is the discovery that a subset of retrotransposons can act as promoters for tissue-specific non-coding RNAs (ncRNAs) or as alternative forms of protein-coding mRNAs[62] (FIG. 3). Discovery of these promoters was made possible through next-generation sequencing, which, unlike microarrays, can distinguish highly related sequences that differ only by a few nucleotides[62]. CAGE studies[63] have identified more than 200,000 human retrotransposon-driven TSSs, which are expressed at low to moderate levels. Frequently, retrotransposon-mediated TSSs start upstream of typical mRNA promoter regions, and they produce RNAs that are transcribed towards the downstream genes.

Transcription factories
Nuclear compartments in which active transcription takes place; they have a high concentration of RNA polymerase II.

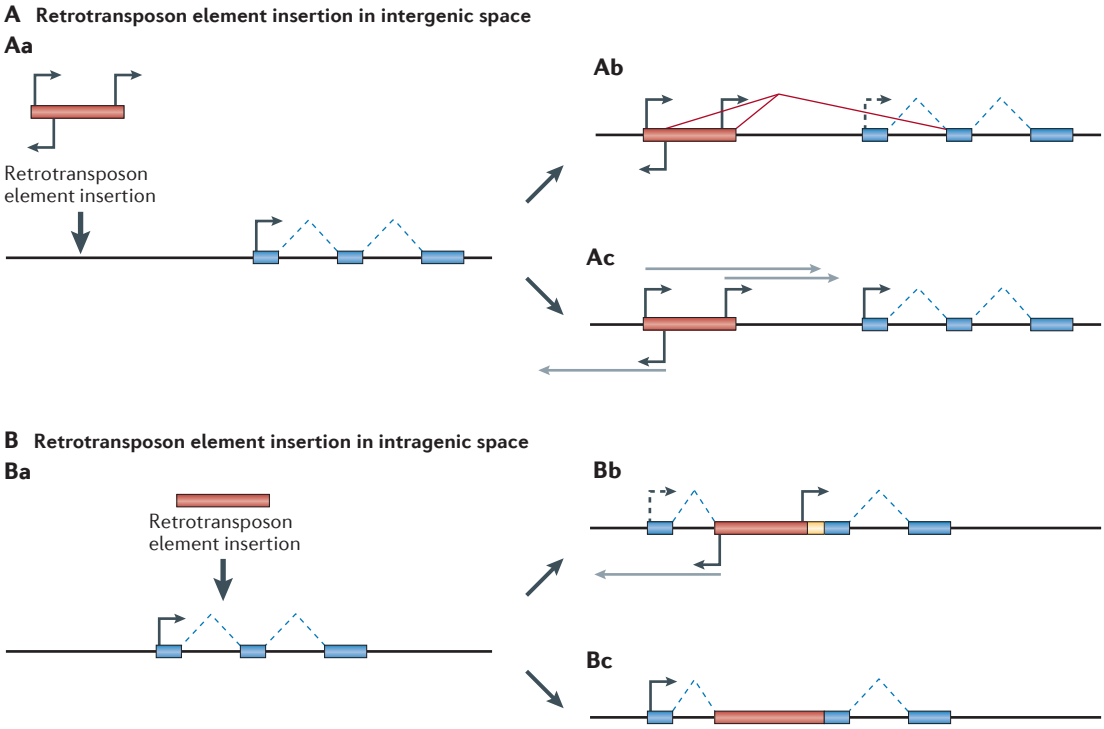


Figure 3 | **Retrotransposon elements influencing transcription. Aa** | A schematic representation of a full-length long interspersed element (LINE) is shown as a red bar with the main transcription sites identified as right-angled arrows. Here it inserts upstream of a gene (exons are shown in blue; splicing is shown as dashed blue lines) in intergenic space. **Ab** | One scenario is that the retrotransposon element acts as an alternative promoter for a nearby gene. Possible splice events are shown as red lines. **Ac** | Alternatively, the retrotransposon element acts as a new promoter that produces non-coding RNAs (grey arrows) that may span other genes or overlap regulatory elements. **Ba** | A retrotransposon element insertion within a gene. **Bb** | This retrotransposon element can provide an internal promoters, driving expression of an alternative form of the gene and/or non-coding RNAs. Yellow represents some intronic sequence that would be included in the mRNA in this scenario. **Bc** | Alternatively, the retrotransposon element can provide an additional exonic sequence.

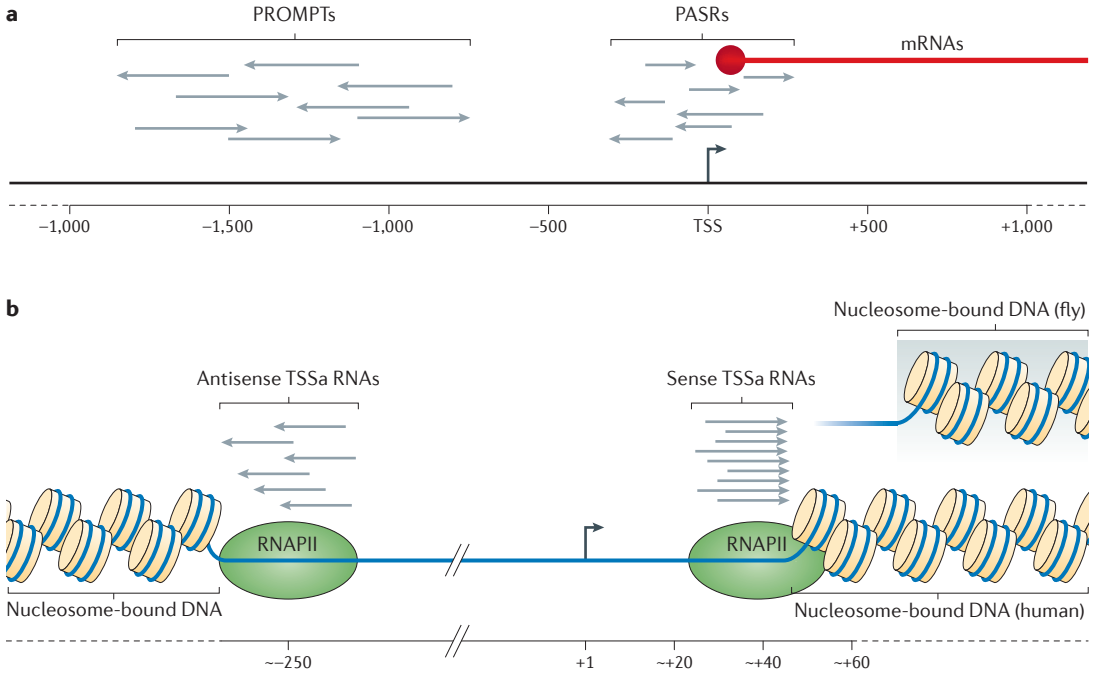
Most of these promoters (except for the highly expressed retrotransposon RNAs) in mammals show low-level transcription from a degenerate pyrimidine–purine initiator consensus that is often reduced to a single G at the position +1 of the RNAs. Often, the CAGE tags start inside the retrotransposon — rather than at the 5′ end — showing that these retrotransposon promoters have complex transcriptional regulation. RNAs that are derived from these promoters often lack the polyadenylation tail[64] and are often localized in the nucleus[65], suggesting that they may have a role in transcriptional regulation and/or nuclear organization. Retrotransposons that are most commonly recruited as promoters tend to be less conserved and incapable of retrotransposition[62]. The emerging idea is that elements that still have the capacity of retrotranspose are generally repressed (for example, by DNA methylation) to reduce the risk of the genomic damage that they might otherwise cause, whereas a subset of elements that are incapable of retrotransposition have been recruited to the regulatory machinery of the genome.

Repeat-derived promoters do not, so far, fit clearly into one of the main promoter classes. Promoters recruited from simple repeats are mostly broad, whereas those that are derived from retrotransposons are sharp but are generally without a TATA box or any other strong spatially constrained motif[62]. With respect to tissue specificity, the few experimentally tested retrotransposon promoters also give mixed results, but they have a preference for tissue-specific activity (reviewed in REF. 66).

Complexity of transcription around gene bodies and 3′ UTR promoters. A surprising finding made using 5′-end-sequencing methods (TABLE 1) was the prevalence of low-intensity aggregates of CAGE tags mapping within exons of protein-coding genes, often on the same strand as the larger gene. Also, there was often a larger aggregate of tags within the 3′ untranslated region (UTR)[8]. The origin and importance of these observations are debated: do they originate from promoters or from degradation of longer RNAs? Some 3′UTR promoters have been validated by reporter assay[8], but recent studies have identified cleavage products of mRNAs, which may be recapped by a cytoplasmic form of the recapping enzyme (reviewed in REFS 67,68), so these products might be the origin of these CAGE tags. An argument for recapping is that CAGE tags are also found to span

Recapping
A process by which an uncapped RNA 5′ end — for example, resulting from degradation — is stabilized by the addition of a cap structure.

Box 1 | **Non-coding RNA species associated with promoters**



A number of non-coding RNA classes occur at or near promoters. These can be roughly divided into longer RNAs that are transcribed in both directions and smaller transcription-start-site-associated (TSSa) RNAs that do not overlap the TSS. The exact number of classes and possible overlap between the classes is under investigation.

Long bidirectional ncRNAs

Promoter upstream transcripts (PROMPTs) are capped bidirectional ncRNAs that map upstream of active promoters (panel **a** of the figure). They are degraded rapidly by the exosome. Their location varies among genes but is typically between −500 and −2,000 with respect to the TSS. They are enriched at CpG island promoters but are present at most expressed genes, including RNA polymerase II (RNAPII) and RNAPIII promoters.

Promoter-associated short RNAs (PASRs) are capped bidirectional RNAs that overlap the start site of mRNAs. They map in the region −200 to +400 in the sense direction and −300 to +200 in the antisense direction with a peak at the TSS. They are 22–100 nt long. Their expression is proportional to that of the gene they overlap, and they tend to track CpG islands.

TSSa RNAs flanking the chromatin-deficient region

TSSa RNAs have a size range of 18–24 nt and are typically found downstream of the TSS (panel **b** of the figure). Their 3′ ends tend to map to approximately +40 on the same strand as the mRNA and to approximately −250 on the antisense strand. Both of these locations are enriched for elongating or poised RNAPII and are adjacent to the first nucleosome in each direction. The expression of these RNAs is associated with CpG islands, and their level of expression correlates with the amount of RNAPII at the promoter. BOX 2 discusses the biogenesis of these RNAs.

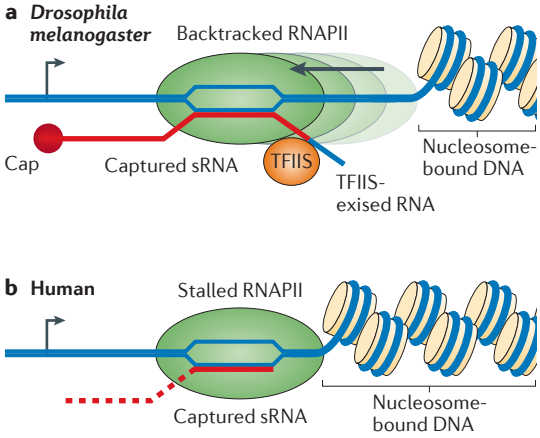
exon–exon junctions and thus originate from mature mRNA. Small RNAs within exons correlated with the incidence of such CAGE tags in the same positions[69], and this association was validated by identifying small RNAs selected with antibodies against the cap structure. Using the complementary PEAT method, it was shown[70] that many such exonic tag aggregates in *D. melanogaster* are indeed capped but lack typical core promoter signals. Similarly, parallel analysis of RNA ends (PARE) was used to identify 5′-monophosphate-cleaved ends of polyadenylated RNA, and a similar association with the CAGE tags was found[71], which would argue for the origin of these tags being degradation and recapping. This study and an earlier one[8] also found that many 5′ ends were tissue- or stage-specific, whereas the correlation with the expression of the annotated full-length mRNA was weak. The latter findings may be reconciled with degradation and recapping only if rapid degradation of the full-length mRNA, together with an efficient stabilization of cleaved RNAs that is perhaps mediated by recapping, occurs. Other models for the origin of the tags have also been proposed based on yeast[72,73] and limited mammalian data[74,75], including the looping of the gene body so that the TSS and termination sites are physically close. In such models, the effective local concentration of RNAPII within 3′UTRs will be higher, and spurious transcriptional initiation will be more likely.

Distinguishing among various components of the signal (biological or technical noise, true signal from promoters or recapping) will be facilitated by using biological replicates and high-throughput methods, such as PARE, ChIP for H3K4me3 or DNAse hypersensitive site analysis, or by using motif analysis (such as enrichment of TATA or Inr motifs) as a computational filter.

Cap structure
A chemical structure found at the 5′ end of mature mRNAs that is used for mRNA stabilization and export to the cytosol.

Box 2 | **Models of biogenesis of promoter-associated small RNAs**

A number of studies have shown small RNAs mapping at or just downstream of transcriptional start sites (TSSs). It is at present unclear whether these are distinct classes: Seila *et al.*[86] defined TSS-associated (TSSa) RNAs as 20–90 nt long RNAs, whereas, in the same regions, Taft *et al.*[87] found smaller, overlapping RNA species (~18 nt on average), which they termed tiny RNAs (tiRNAs). These tiRNAs end at approximately +40. Two models for their biogenesis, which are both associated with RNA polymerase II (RNAPII) properties, have been proposed. These two models are not mutually exclusive and could collectively account for the mixture of capped and uncapped small RNAs of diverse lengths that are observed.



Backtracking and excision

In this model[59,60,88], RNA polymerase II (RNAPII) backtracks towards the TSS, which exposes a region of the nascent RNA. This region is cleaved by TFIIS (also known as TCEA1). A problem with this model is that the tiRNA size range (>18 nt) is not compatible with the size of RNAs that *in vitro* experiments have determined to be excised (6–14 nt)[89]. However, in flies (panel **a** of the figure), there is another subclass of small RNAs that are capped, that start around the real mRNA TSS and that terminate at +38. These small RNAs extend further downstream if TFIIS is depleted, indicating that this subgroup of RNAs is the product of backtracking and TFIIS-mediated cleavage[60]. This is consistent with the nucleosome boundary, which in flies is ~20 nt downstream of the boundary in mammals.

Unsuccessful RNAPII elongation followed by degradation

A subset of RNAPII complexes initiate transcription but do not undergo elongation. Nascent RNAs (which are not capped) will then be degraded from the 5′ end, but the part of the RNA that is covered by RNAPII will be protected from the degradation enzymes (panel **b** of the figure). RNAPII will cover about 17–22 nt of the RNA, which fits with the size range of tiRNAs. So, in this model, the tiRNAs are the result of RNA degradation and RNAPII protection. This model also suggests that RNAPII stalling may have a role as an RNA-capping quality checkpoint[81,90].

Promoter-associated small RNAs. A series of studies has identified different families of ncRNAs around promoters of protein-coding genes (BOX 1). These have different size ranges, chemical modifications and suggested modes of biogenesis, but they can roughly be divided by their size range and location respective to TSSs. A first broad class encompasses larger RNAs (>100 nt) that can be transcribed from the same region in both directions. These include promoter-associated long RNAs (PALRs) and promoter-associated short RNAs (PASRs), which are identified by tiling arrays[76] (BOX 1a). Parts of these transcripts overlap with a currently poorly understood class of unstable ncRNAs called promoter upstream transcripts (PROMPTs), which are located from ~0.5 to 2 kb upstream of TSSs or of known mRNAs[77]. Possible functions for these RNAs include being precursors for small regulatory RNAs or interacting with chromatin[76].

A second set comprises ncRNAs that are small and that originate at the TSS or just downstream of it. They are transcribed in the same direction as the protein-coding genes and also in the reverse direction 250 nt upstream of the TSS (BOX 1, also reviewed in REF. 78 and in REF. 79). The upstream and downstream small RNA peaks correlate with the RNAPII ChIP signals, and they are also located near the edge of the nucleosomes immediately upstream and downstream of the TSS[80] (BOX 1b). This suggests that the biogenesis of these RNAs is linked to RNAPII during stalling and/or during early elongation (BOX 2). The presence of these small RNAs seems to be a general feature of active promoters and correlates with the amount of RNAPII, but the relative level of the RNAs up- and downstream of the TSS are related to signals in the promoter, linking it to the classes of promoters discussed above. Sharp (type I) promoters tend to have substantially more small RNAs downstream of the TSS, whereas broad promoters (mostly type II) have a more even distribution[81]. The likely explanation for this is that the small RNAs are by-products of elongating RNAPII, and sharp promoters provide much more directionality for RNAPII owing to the presence of TATA boxes and similar elements.

We face a number of challenges relating to these small RNAs. First, it is not clear how many distinct classes exist, as the studies reporting them have used varied methodologies — thus, many of the classes that are suggested to be distinct might actually have the same biogenesis. Second, although many current studies indicate that many of these small RNAs are by-products of transcription and are therefore expected to have a limited function, there are notable exceptions that demonstrate function[82]; strategies to separate these are necessary. Finally, further work is needed to relate small RNAs to classes of promoters.

Perspective: dimensions of transcriptional regulation

Our knowledge of the structural and functional properties of promoters has made enormous advances in recent years. Here we have discussed how separate threads of information about promoter structure, function and

classification are starting to come together, primarily thanks to an enormous amount of genome-wide data gathered by high-throughput-sequencing-based methods. Different functional roles of genes and the need to control their dynamics in fundamentally different ways have not only resulted in different regulatory elements but also in promoter architectures that interpret these regulatory inputs in different ways. For example, promoters controlling developmental genes require different responses to the same regulatory factors than promoters at genes that are expressed in one specific context. This distinction was apparently established early in metazoan evolution, although the distinction between ubiquitously expressed and regulated genes possibly appeared even earlier[83,84]. In metazoans with large genomes, at least part of the expanded sequence has been populated by clusters of regulatory elements targeting key genes that control complex multicellular developmental processes[42,85]. The ability of promoters to respond to such a complex array of inputs is one of the principal processes that deserve our attention in future studies.

Only a minority of promoters fit the 'classical' model of transcriptional initiation and regulation: tissue-specific sharp or peaked (type I) promoters, which have most of their regulatory elements close to the TSS and are controlled locally. A much larger fraction of genes is regulated by broad promoters with activity that seems to be more influenced by the epigenomic context and less so by sequence-specific transcription factors. As developmental genes and ubiquitously expressed genes need to be switched on and off under distinct circumstances, these different promoter classes are clearly using different aspects of the cellular repertoire of regulatory mechanisms that work over a range of genomic distances. Another type of regulatory entity that needs to be fitted with the promoter classes is ncRNAs.

An inherent challenge with almost all methods discussed in this Review is data analysis, as a single experiment often gives 20–70 million DNA reads. This is not solely a technical challenge that can be addressed by sheer computer power; it also requires substantial human expertise to go from raw data to biologically testable hypotheses. We believe that large-scale projects such as ENCODE and FANTOM are highly beneficial not only owing to their data production but also owing to their unification of analysis methods, use of common cell lines and their larger body of expertise in data interpretation and modelling than in any single research group. As a community, we have an added challenge of educating the next generation of researchers to be able to handle the challenges of both experimental and computational analysis.

1. Sandelin, A. *et al.* Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nature Rev. Genet.* **8**, 424–436 (2007).
2. Valen, E. & Sandelin, A. Genomic and chromatin signals underlying transcription start-site selection. *Trends Genet.* **27**, 475–485 (2011).
3. Maston, G. A., Evans, S. K. & Green, M. R. Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.* **7**, 29–59 (2006).
4. Riethoven, J.-J. M. Regulatory regions in DNA: promoters, enhancers, silencers, and insulators. *Methods Mol. Biol.* **674**, 33–42 (2010).
5. Ohler, U. & Wassarman, D. A. Promoting developmental transcription. *Development* **137**, 15–26 (2010).
6. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **25**, 1010–1022 (2011).
7. Kadonaga, J. T. Perspectives on the RNA polymerase II core promoter. *WIREs Dev. Biol.* **1**, 40–51 (2012).
8. Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genet.* **38**, 626–635 (2006).
This is one of the most comprehensive early studies on TSS distributions in humans and mice.
9. Yamashita, R., Suzuki, Y., Sugano, S. & Nakai, K. Genome-wide analysis reveals strong correlation between CpG islands with nearby transcription start sites of genes and their tissue specificity. *Gene* **350**, 129–136 (2005).
10. Yoshimura, K. *et al.* The cystic fibrosis gene has a "housekeeping"-type promoter and is expressed at low levels in cells of epithelial origin. *J. Biol. Chem.* **266**, 9140–9144 (1991).
11. Ponjavic, J. *et al.* Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters. *Genome Biol.* **7**, R78 (2006).
12. Plessy, C. *et al.* Promoter architecture of mouse olfactory receptor genes. *Genome Res.* 22 Dec 2011 (doi:10.1101/gr.126201.111).
13. Rach, E. A. *et al.* Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level. *PLoS Genet.* **7**, e1001274 (2011).
This is a study that correlated TSS shapes with chromatin mark information, showing the link between the two features.
14. FitzGerald, P. C., Sturgill, D., Shyakhtenko, A., Oliver, B. & Vinson, C. Comparative genomics of *Drosophila* and human core promoters. *Genome Biol.* **7**, R53 (2006).
15. Ohler, U. Identification of core promoter modules in *Drosophila* and their application in accurate transcription start site prediction. *Nucleic Acids Res.* **34**, 5943–5950 (2006).
16. Engstrom, P. G., Ho Sui, S. J., Drivenes, O., Becker, T. S. & Lenhard, B. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome Res.* **17**, 1898–1908 (2007).
17. Akalin, A. *et al.* Transcriptional features of genomic regulatory blocks. *Genome Biol.* **10**, R38 (2009).
18. Rach, E. A., Yuan, H.-Y., Majoros, W. H., Tomancak, P. & Ohler, U. Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the *Drosophila* genome. *Genome Biol.* **10**, R73 (2009).
19. Hoskins, R. A. *et al.* Genome-wide analysis of promoter architecture in *Drosophila melanogaster*. *Genome Res.* **21**, 182–185 (2011).
20. Hendrix, D. A., Hong, J.-W., Zeitlinger, J., Rokhsar, D. S. & Levine, M. S. Promoter elements associated with RNA Pol II stalling in the *Drosophila* embryo. *Proc. Natl Acad. Sci. USA* **105**, 7762–7767 (2008).
21. van Heeringen, S. J. *et al.* Nucleotide composition-linked divergence of vertebrate core promoter architecture. *Genome Res.* **21**, 410–421 (2011).
22. Grishkevich, V., Hashimshony, T. & Yanai, I. Core promoter T-blocks correlate with gene expression levels in *C. elegans*. *Genome Res.* **21**, 707–717 (2011).
23. Parry, T. J. *et al.* The TCT motif, a key component of an RNA polymerase II transcription system for the translational machinery. *Genes Dev.* **24**, 2013–2018 (2010).
24. Damgaard, C. K. & Lykke-Andersen, J. Translational coregulation of 5′TOP mRNAs by TIA-1 and TIAR. *Genes Dev.* **25**, 2057–2068 (2011).
25. Perry, R. P. The architecture of mammalian ribosomal protein promoters. *BMC Evol. Biol.* **5**, 15 (2005).
26. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotech.* **28**, 817–825 (2010).
27. Zeitlinger, J. *et al.* RNA polymerase stalling at developmental control genes in the *Drosophila melanogaster* embryo. *Nature Genet.* **39**, 1512–1516 (2007).
This is one of several papers that used genomics methods to decipher stalling or poising; it also revealed functional tripartition of promoters based on RNAPII signal.
28. Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315–326 (2006).
29. Schwartz, Y. B. *et al.* Alternative epigenetic chromatin states of Polycomb target genes. *PLoS Genet.* **6**, e1000805 (2010).
30. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
This study uses an algorithm to segment the genome of nine ENCODE cell lines into regions with different functions based on the combination of epigenetic marks, revealing genome-wide epigenetic differences between promoter classes.
31. Kharchenko, P. V. *et al.* Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. *Nature* **471**, 480–485 (2011).
This paper discusses a genome-wide chromatin landscape for *D. melanogaster* based on comprehensive histone modifications identifying combinatorial patterns, which is further integrated with chromosomes, genes and regulatory elements characteristics.
32. Izzo, A. & Schneider, R. Chatting histone modifications in mammals. *Brief Funct. Genomics* **9**, 429–443 (2010).
33. Lee, J.-S., Smith, E. & Shilatifard, A. The language of histone crosstalk. *Cell* **142**, 682–685 (2010).
34. Nozaki, T. *et al.* Tight associations between transcription promoter type and epigenetic variation in histone positioning and modification. *BMC Genomics* **12**, 416 (2011).
35. Jiang, C. & Pugh, B. Nucleosome positioning and gene regulation: advances through genomics. *Nature Rev. Genet.* **10**, 161–172 (2009).
36. Ioshikhes, I., Hosid, S. & Pugh, F. Variety of genomic DNA patterns for nucleosome positioning. *Genome Res.* **21**, 1863–1871 (2011).
37. Radman-Livaja, M., Liu, C. L., Friedman, N., Schreiber, S. L. & Rando, O. J. Replication and active demethylation represent partially overlapping mechanisms for erasure of H3K4me3 in budding yeast. *PLoS Genet.* **6**, e1000837 (2010).
38. Ramirez-Carrozzi, V. R. *et al.* A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling. *Cell* **138**, 114–128 (2009).

39. Subtil-Rodríguez, A. & Reyes, J. C. BRG1 helps RNA polymerase II to overcome a nucleosomal barrier during elongation, *in vivo*. *EMBO Rep.* **11**, 751–757 (2010).
40. Hargreaves, D. C., Horng, T. & Medzhitov, R. Control of inducible gene expression by signal-dependent transcriptional elongation. *Cell* **138**, 129–145 (2009).
41. Fu, Y., Sinha, M., Peterson, C. L., Weng, Z. & van Steensel, B. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet.* **4**, e1000138 (2008).
42. Kikuta, H. *et al.* Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res.* **17**, 545–555 (2007).
43. Visel, A., Rubin, E. M. & Pennacchio, L. A. Genomic views of distant-acting enhancers. *Nature* **461**, 199–205 (2009).
44. Mikkelsen, T. S. *et al.* Comparative epigenomic analysis of murine and human adipogenesis. *Cell* **143**, 156–169 (2010).
45. Roider, H. G., Lenhard, B., Kanhere, A., Haas, S. A. & Vingron, M. CpG-depleted promoters harbor tissue-specific transcription factor binding signals—implications for motif overrepresentation analyses. *Nucleic Acids Res.* **37**, 6305–6315 (2009).
46. Soler, E. *et al.* A systems approach to analyze transcription factors in mammalian cells. *Methods* **53**, 151–162 (2011).
47. Dean, A. In the loop: long range chromatin interactions and gene regulation. *Brief Funct. Genomics* **10**, 3–10 (2011).
48. Cremer, T. & Cremer, M. Chromosome territories. *Cold Spring Harb. Perspect. Biol.* **2**, a003889 (2010).
49. Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948–951 (2008).
50. Lanctôt, C., Cheutin, T., Cremer, M., Cavalli, G. & Cremer, T. Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nature Rev. Genet.* **8**, 104–115 (2007).
51. Ferrai, C., de Castro, I. J., Lavitas, L., Chotalia, M. & Pombo, A. Gene positioning. *Cold Spring Harb. Perspect. Biol.* **2**, a000588 (2010).
52. Muse, G. W. *et al.* RNA polymerase is poised for activation across the genome. *Nature Genet.* **39**, 1507–1511 (2007).
53. Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R. & Young, R. A. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130**, 77–88 (2007).
This was one of several papers using genomics methods to decipher stalling or poising.
54. Nechaev, S. & Adelman, K. *Pol.* II waiting in the starting gates: regulating the transition from transcription initiation into productive elongation. *Biochim. Biophys. Acta* **1809**, 34–45 (2011).
55. Gilmour, D. S. & Lis, J. T. RNA polymerase II interacts with the promoter region of the noninduced *hsp70* gene in *Drosophila melanogaster* cells. *Mol. Cell. Biol.* **6**, 3984–3989 (1986).
56. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA Sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845–1848 (2008).
57. Buratowski, S. Progression through the RNA polymerase II CTD Cycle. *Mol. Cell* **36**, 541–546 (2009).
58. Ferrai, C. *et al.* Poised transcription factories prime silent uPA gene prior to activation. *PLoS Biol.* **8**, e1000270 (2010).
59. Shaevitz, J. W., Abbondanzieri, E. A., Landick, R. & Block, S. M. Backtracking by single RNA polymerase molecules observed at near-base-pair resolution. *Nature* **426**, 684–687 (2003).
60. Nechaev, S. *et al.* Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in *Drosophila*. *Science* **327**, 335–338 (2010).
61. Gilchrist, D. A. *et al.* Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation. *Cell* **143**, 540–551 (2010).
62. Faulkner, G. J. *et al.* The regulated retrotransposon transcriptome of mammalian cells. *Nature Genet.* **41**, 563–571 (2009).
This paper showed the large number of retrotransposon elements that are potential TSSs.
63. Frith, M. C. *et al.* A code for transcription initiation in mammalian genomes. *Genome Res.* **18**, 1–12 (2008).
64. Faulkner, G. J. & Carninci, P. Altruistic functions for selfish DNA. *Cell Cycle* **8**, 2895–2900 (2009).
65. Plessy, C. *et al.* Linking promoters to functional transcripts in small samples with nanoCAGE and CAGEscan. *Nature Methods* **7**, 528–534 (2010).
66. Cohen, C. J., Lock, W. M. & Mager, D. L. Endogenous retroviral LTRs as promoters for human genes: a critical assessment. *Gene* **448**, 105–114 (2009).
67. Schoenberg, D. R. & Maquat, L. E. Re-capping the message. *Trends Biochem. Sci.* **34**, 435–442 (2009).
68. Jackowiak, P., Nowacka, M., Strozycki, P. M. & Figlerowicz, M. RNA degradome—its biogenesis and functions. *Nucleic Acids Res.* **39**, 7361–7370 (2011).
69. Fejes-Toth, K. *et al.* Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. *Nature* **457**, (2009) (1028).
70. Ni, T. *et al.* A paired-end sequencing strategy to map the complex landscape of transcription initiation. *Nature Methods* **7**, 521–527 (2010).
71. Mercer, T. R. *et al.* Regulated post-transcriptional RNA cleavage diversifies the eukaryotic transcriptome. *Genome Res.* **20**, 1639–1650 (2010).
72. O'Sullivan, J. M. *et al.* Gene loops juxtapose promoters and terminators in yeast. *Nature Genet.* **36**, 1014–1018 (2004).
73. Kaderi, El, B., Medler, S., Raghunayakula, S. & Ansari, A. Gene looping is conferred by activator-dependent interaction of transcription initiation and termination machineries. *J. Biol. Chem.* **284**, 25015–25025 (2009).
74. Perkins, K. J., Lusic, M., Mitar, I., Giacca, M. & Proudfoot, N. J. Transcription-dependent gene looping of the HIV-1 provirus is dictated by recognition of pre-mRNA processing signals. *Mol. Cell* **29**, 56–68 (2008).
75. Tan-Wong, S. M., French, J. D., Proudfoot, N. J. & Brown, M. A. Dynamic interactions between the promoter and terminator regions of the mammalian *BRCA1* gene. *Proc. Natl Acad. Sci. USA* **105**, 5160–5165 (2008).
76. Kapranov, P. *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**, 1484–1488 (2007).
77. Preker, P. *et al.* PROMoter uPstream Transcripts share characteristics with mRNAs and are produced upstream of all three major types of mammalian promoters. *Nucleic Acids Res.* **39**, 7179–7193 (2011).
78. Carninci, P. RNA dust: where are the genes? *DNA Res.* **17**, 51–59 (2010).
79. Jacquier, A. The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nature Rev. Genet.* **10**, 833–844 (2009).
80. Taft, R. J., Kaplan, C. D., Simons, C. & Mattick, J. S. Evolution, biogenesis and function of promoter-associated RNAs. *Cell Cycle* **8**, 2332–2338 (2009).
81. Valen, E. *et al.* Biogenic mechanisms and utilization of small RNAs derived from human protein-coding genes. *Nature Struct. Mol. Biol.* **18**, 1075–1082 (2011).
82. Cernilogar, F. M. *et al.* Chromatin-associated RNA interference components contribute to transcriptional regulation in *Drosophila*. *Nature* **480**, 391–395 (2011).
83. Basehoar, A. D., Zanton, S. J. & Pugh, B. F. Identification and distinct regulation of yeast TATA box-containing genes. *Cell* **116**, 699–709 (2004).
84. Yamamoto, Y. Y. *et al.* Heterogeneity of *Arabidopsis* core promoters revealed by high-density TSS analysis. *Plant J.* **60**, 350–362 (2009).
85. Woolfe, A. *et al.* Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* **3**, e7 (2005).
86. Seila, A. C. *et al.* Divergent transcription from active promoters. *Science* **322**, 1849–1851 (2008).
87. Taft, R. J. *et al.* Tiny RNAs associated with transcription start sites in animals. *Nature Genet.* **41**, 572–578 (2009).
88. Taft, R. J. *et al.* Nuclear-localized tiny RNAs are associated with transcription initiation and splice sites in metazoans. *Nature Struct. Mol. Biol.* **17**, 1030–1034 (2010).
89. Izban, M. G. & Luse, D. S. The increment of SII-facilitated transcript cleavage varies dramatically between elongation competent and incompetent RNA polymerase II ternary complexes. *J. Biol. Chem.* **268**, 12874–12885 (1993).
90. Mandal, S. S. *et al.* Functional interactions of RNA-capping enzyme with factors that positively and negatively regulate promoter escape by RNA polymerase II. *Proc. Natl Acad. Sci. USA* **101**, 7572–7577 (2004).
91. Wasserman, W. W. & Sandelin, A. Applied bioinformatics for the identification of regulatory elements. *Nature Rev. Genet.* **5**, 276–287 (2004).
92. Valen, E. *et al.* Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. *Genome Res.* **19**, 255–265 (2009).
93. Kanamori-Katayama, M. *et al.* Unamplified cap analysis of gene expression on a single-molecule sequencer. *Genome Res.* **21**, 1150–1159 (2011).
94. Affymetrix/Cold Spring Harbor Laboratory ENCODE Transcriptome Project. Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. *Nature* **457**, 1028–1032 (2009).
This study shows the large diversity of ncRNAs around promoters.
95. Hashimoto, S.-I. *et al.* 5′-end SAGE for the analysis of transcriptional start sites. *Nature Biotech.* **22**, 1146–1149 (2004).
96. Ng, P. *et al.* Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nature Methods* **2**, 105–111 (2005).
97. Thomas, M. F. & Ansel, K. M. Construction of small RNA cDNA libraries for deep sequencing. *Methods Mol. Biol.* **667**, 93–111 (2010).
98. Kawaji, H. *et al.* Hidden layers of human small RNAs. *BMC Genomics* **9**, 157 (2008).
99. Landgraf, P. *et al.* A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* **129**, 1401–1414 (2007).
100. The ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* **9**, e1001046 (2011).
101. Gilchrist, D. A. *et al.* NELF-mediated stalling of Pol II can enhance gene expression by blocking promoter-proximal nucleosome assembly. *Genes Dev.* **22**, 1921–1933 (2008).
102. Gries, T. J., Kontur, W. S., Capp, M. W., Saecker, R. M. & Record, M. T. One-step DNA melting in the RNA polymerase cleft opens the initiation bubble to form an unstable open complex. *Proc. Natl Acad. Sci. USA* **107**, 10418–10423 (2010).

Acknowledgements
B.L. acknowledges the support of the Bergen Research Foundation, the Norwegian YFF project 180435, the Norwegian Research Foundation and the UK Medical Research Council. A.S. was supported by grants from the European Research Commission (FP7/2007-2013/ERC grant agreement 204135), The Novo Nordisk Foundation, The Lundbeck Foundation and the Danish Cancer Society. P.C. was supported by a grant from the Seventh Framework of the European Union Commission to the Dopaminet Consortium, the Modhep Consortium, the Braintrain Consortium, the Funding Program for the Next Generation World-Leading Researchers (NEXT Program) and a research grant to RIKEN Omics Science Center from the Japanese Ministry of Education, Culture, Sports, Science and Technology.

Competing interests statement
The authors declare competing financial interests: see Web version for details.

FURTHER INFORMATION
Boris Lenhard's homepage: http://www.csc.mrc.ac.uk/research/groups/compgen
Albin Sandelin's homepage: http://people.binf.ku.dk/albin/Sandelin_group_at_the_Bioinformatic_Centre/The_Sandelin_group.html
Piero Carninci's homepage: http://genome.gsc.riken.jp/osc/english/members/Piero_Carninci.html
ENCODE: http://www.genome.gov/10005107
FANTOM: http://fantom.gsc.riken.jp/4
modENCODE: http://www.modencode.org
***Nature Reviews Genetics Series* on Modes of transcriptional regulation:** http://www.nature.com/nrg/series/transcriptionalregulation/index.html
***Nature Reviews Genetics Series* on Regulatory elements:** http://www.nature.com/nrg/series/regulatoryelements/index.html
ALL LINKS ARE ACTIVE IN THE ONLINE PDF