Promoter Architectures and Developmental Gene Regulation

Vanja Haberle[a,b,1] and Boris Lenhard[a,*]

[a] Institute of Clinical Sciences and MRC Clinical Sciences Center, Faculty of Medicine, Imperial College London, Hammersmith Hospital, Du Cane Road, London W12 0NN, UK.

[b] Department of Biology, University of Bergen, Thormøhlensgate 53A, N-5008 Bergen, Norway.

[*] Corresponding author. E-mail address: b.lenhard@imperial.ac.uk (B. Lenhard)

[1] Present address: Research Institute of Molecular Pathology (IMP), Vienna Biocenter (VBC), Dr Bohr-Gasse 7, 1030 Vienna, Austria.

Abstract

Core promoters are minimal regions sufficient to direct accurate initiation of transcription and are crucial for regulation of gene expression. They are highly diverse in terms of associated core promoter motifs, underlying sequence composition and patterns of transcription initiation. Distinctive features of promoters are also seen at the chromatin level, including nucleosome positioning patterns and presence of specific histone modifications. Recent advances in identifying and characterizing promoters using next-generation sequencing-based technologies have provided the basis for their classification into functional groups and have shed light on their modes of regulation, with important implications for transcriptional regulation in development. This review discusses the methodology and the results of genome-wide studies that provided insight into the diversity of RNA polymerase II promoter architectures in vertebrates and other Metazoa, and the association of these architectures with distinct modes of regulation in embryonic development and differentiation.

Keywords

transcriptional regulation; core promoter; transcription start sites; CAGE; promoter types; overlapping codes

1. Transcriptional machinery and RNAPII core promoters

Protein-coding genes and several classes of noncoding RNA (ncRNA) genes are transcribed by RNA polymerase II (RNAPII), a large multi-subunit enzyme that uses DNA as a template to produce complementary RNA molecule [1]. RNAPII initiates transcription at individual nucleotides at the beginning of the gene called transcription start sites (TSS). The region surrounding TSS is known as core promoter and it is defined as a minimal region that is sufficient to direct the accurate initiation of transcription. Core promoter typically extends ~40 bp upstream and downstream of the TSS, and it is a place of the assembly of the transcriptional machinery [2]. This process requires general transcription factors (GTF), which recognize and bind core promoter elements and recruit RNAPII. There are six general transcription factors: TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH, which assemble to core promoter in a stepwise manner and form a pre-initiation complex (PIC) [1]. TFIID plays a central role in recognising and binding specific core promoter elements and creates an environment that facilitates transcription initiation [3].

Various core promoter elements have been identified in eukaryotic promoters and include a TATA-box, an Initiator (Inr), a Downstream Promoter Element (DPE), a Downstream Core Element (DCE), a TFIIB-Recognition Element (BRE), and a Motif Ten Element (MTE) [4] (Fig. 1). However, none of these elements are universal, since they are found only in a fraction of core promoters in various combinations and there are many promoters that lack any of those elements [5]. In addition, some core promoter elements are associated with specific biological functions, for instance the TCT motif, which is found exclusively in promoters of genes that encode the components of the translational machinery [6].

Many core promoters in vertebrates overlap with CpG islands (CGI), which are genomic regions characterised by elevated C+G content and frequency of the CG dinucleotides compared to the bulk genome [7,8]. The current estimation is that ~70% of human promoters are associated with a CGI [9], with similar percentage observed for mouse and chicken [10]. The proportion of CpG

promoters seems substantially lower in amphibians and fish [10]. However, this is likely due to the fact that the definition of the CpG island relies on arbitrary thresholds set upon C+G content, observed over expected ratio of CpG dinucleotide counts and region length [11], which have been optimised for mammalian genomes and do not perform well for genomes with very different nucleotide composition. Nevertheless, association with CpG islands distinguishes two main classes of vertebrate promoters, high-CpG promoters and low-CpG promoters [9,12], which are additionally characterised by distinct promoter features and functions of associated genes [12].

The complexity of the core promoter is further seen in the relation among specificity of expression, transcription initiation patterns, motif composition and the organization of the chromatin structure in the promoter region, as discussed further below. All this suggests that core promoters are not passive elements that serve only to direct the proper placement of the RNA polymerase II transcriptional machinery. They receive and integrate various regulatory inputs and convert them into precise rate of transcription initiation. Core promoter elements can determine the responsiveness of the promoter to transcriptional regulation by *cis*-regulatory elements and *trans*-acting factors in multicellular organisms [13,14], and are major determinants of gene expression level in yeast [15], making them central, active components of transcriptional regulation.

2. Single-nucleotide transcription initiation data is central to studying promoter architecture

Mapping promoters genome-wide is the first step in deciphering the mechanisms of transcriptional regulation and different approaches have been used to detect promoters along the genome experimentally. Various features of active promoters such as the presence of the PIC, promoter-associated histone marks and accessible and open chromatin have been used to localise promoters [16]. These approaches can only identify loci that serve as promoters, but cannot map precise transcription start sites or quantify the level of transcription from the detected promoters. Since transcriptionally active promoters produce transcripts, an alternative approach is to use the expression data to derive positions of the promoters. However, majority of the transcriptomic data maps transcribed portions of the genome but does not precisely reflect gene boundaries. For instance, typical expressed sequence tag represents only a random short subsequence of the full cDNA. Furthermore, RNA-seq, which is the most common technique for quantitative

transcriptome profiling, produces uneven coverage of sequenced tags along the transcript, often not covering the 5' end [17]. In order to precisely map promoters, 5' end complete cDNAs are essential. First genome-wide sequencing and annotation of full-length cDNAs was done for mouse by the FANTOM Consortium [18] and used to determine exact TSSs and characterise adjacent putative promoter regions. Similarly, full-length human cDNAs were used to annotate and functionally analyse human promoters [19,20]. More recently, several techniques that sequence short tags from the 5' end of cDNAs have been developed including 5' serial analysis of gene expression (5' SAGE) [21], oligo-capping [22] and cap analysis of gene expression (CAGE) [23], which when combined with high-throughput sequencing achieve higher coverage producing more reliable and quantitative mapping of 5' ends. These techniques allow genome-wide precise TSS mapping at single nucleotide resolution and provide the means for analysing promoter-associated features at high resolution.

Precise, 1bp resolution mapping of transcription start sites has proven to be central to studying the details of promoter architectures and for their classification. Even though most of the strong promoter motifs and their respective locations were estimated from relatively modest amounts of pre-genome data, CAGE has provided evidence that the distance between TSS and the motifs such as TATA box or DPE is much more constrained than thought previously [24,25] – indeed, promoters with single well defined TSS are usually characterised by a fixed spacing from a motif that defines them. In addition, even weaker motifs such as nucleosome positioning sequence and general GC composition have been shown to line up precisely with most commonly used TSS position in promoters with broader initiation pattern [26,27], revealing hitherto unknown global features of this type of promoter.

3. Methodologies for precise transcription start site identification

3.1. Cap Analysis of Gene Expression (CAGE)

CAGE is a high-throughput method for transcriptome analysis [23] that takes advantage of the 7-methylguanosine cap structure found at 5' ends of RNAPII transcripts to map precise transcription start sites (Fig. 2). The protocol includes biotinylation of the cap structure, reverse transcription, and treatment of the RNA/DNA heteroduplex with RNase I to ensure that only 5'-complete cDNAs stay associated with the biotin tag and are pulled down by streptavidin-coated beads. A linker sequence containing recognition site for type III restriction endonuclease is ligated

to the 5' end of the captured cDNA and a corresponding restriction enzyme is used to cleave off a short fragment (typically 27 bp) from the 5' end [28]. The resulting fragments are then amplified and sequenced using massive parallel high-throughput sequencing technology, which results in a large number of short sequenced tags that can be mapped back to the reference genome to infer the exact position of the TSSs used to initiate transcription of captured RNAs. Number of CAGE tags supporting each CAGE-detected TSS (CTSS) gives the information on the relative frequency of its usage and can be used as a measure of expression from that specific TSS [29]. Thus, CAGE provides information on two aspects of the capped transcriptome: 1) genome-wide single base-pair resolution map of transcription start sites, and 2) relative levels of transcripts initiated at each CTSS. This information can be used for various analyses, from 5' end centred expression profiling [30,31] to studying promoter architecture [12,25].

Since the introduction of CAGE, a great effort has been made by the FANTOM consortium to map genome-wide TSSs in numerous mouse and human samples [27,32,33]. This has led to the discovery of distinct classes of promoters with respect to TSS distribution that correlates with both underlying sequence features and gene function [12], and implies distinct modes of their regulation (reviewed in [34]). Quantitative nature of CAGE has been used to model expression dynamics and to reconstruct the regulatory networks driving the differentiation [30] and maintaining identity of numerous human and mouse cell and tissue types [27], by identifying key transcription factors binding at promoters. Moreover, CAGE signal has been shown to be enriched at enhancers [35] and has been used to construct an atlas of active enhancers over cells and tissues across the whole human body [36]. Thus, in addition to providing a valuable resource of genome-wide cell type-specific TSSs, which are a more precise alternative to TSS positions available in annotation databases, CAGE is also a powerful approach for studying various aspects of gene regulation.

However, not all genomic positions detected by CAGE seem to correspond to genuine RNAPII transcription initiation sites, as many CTSSs were found within internal exons with CAGE tags spanning exon-exon junctions [12]. A study profiling small RNAs and comparing them to distribution of CAGE tags concluded that processed coding and non-coding RNAs are metabolized into short RNAs that likely bear cap-like structures at their 5’ ends and are captured by CAGE tags [37]. The function of these short and CAGE-sensitive RNAs mapping to internal exons and introns remains elusive. However, these RNA species arise only from a discrete subset of genes and their abundance often does not correlate with the expression of the host gene, arguing against them being merely degradation intermediates [37,38].

3.2. Mapping TSSs of nascent transcripts

One limitation of the CAGE protocol is that it works on total mature RNA or a specific fraction thereof (such as polyA RNA transcripts, or RNA isolated from specific cellular compartments). In practice, this means that the CAGE TSS signal does not reflect the state of TSS usage at the time of RNA isolation, but rather the 5' ends of transcripts that have accumulated in an undefined time window prior to the isolation. This enriches the signal for the TSS of long-lived transcripts. It also introduces delay and decreases temporal resolution in time-course experiments. To overcome these limitations, novel approaches for detecting nascent transcripts (GRO-seq) have been coupled with techniques for capturing the 5' cap structure, and have been recently used to map TSSs of nascent transcripts at single base-pair resolution [39]. By mapping both stable and unstable RNAs, the GRO-cap approach has revealed the precise architecture of pervasive divergent transcription initiation in human genome, as well as its underlying sequence and chromatin features [39].

4. Chromatin structure, modifications and epigenetic data aid in genome-wide analysis of promoter architectures

4.1. Nucleosome positioning at promoters

Genetic information is encoded in DNA in a linear fashion. However, to enable efficient storage, organisation and control of large amount of DNA within the nucleus, the linear DNA molecules are coupled with histone and other non-histone proteins into a macromolecular complex known as chromatin. Histone octamers bound by approximately 147 bp of DNA form nucleosomes, which are arranged as a linear array along the DNA polymer creating a "beads on a string" structure. The packaging of DNA creates both a problem and an opportunity, since wrapping of DNA around histones potentially obstructs access to the genetic code. However, the ubiquity of nucleosomes at all regions of chromosomal DNA can be exploited to direct the enzymes that read, replicate and repair DNA to the appropriate entry sites.

Nucleosome positioning was most extensively studied in the compact yeast genome, and the first genome-wide mapping of nucleosome positions at high resolution showed that the nucleosomes at most genes are generally organized in the same way [40]. Around the beginning of a gene there is a nucleosome free region (NFR) flanked by two well-positioned nucleosomes (the –1 and +1

nucleosomes), which is followed by an array of nucleosomes that package the gene body. The first, +1 nucleosome, displays the tightest positioning and is subject to various histone protein variants and modifications, implicating its involvement in regulation of gene transcription. Further downstream nucleosomes exhibit lower levels of phasing. This basic pattern was later shown to be present in metazoan genomes as well [41,42].

In contrast, the vast majority of nucleosomes throughout the rest of the genome seem to be positioned with expected periodicity and form arrays of phased nucleosomes around barriers imposed by DNA binding proteins or minority of well-positioned nucleosomes [43]. Despite the controversy around the degree to which primary sequence determines nucleosome positioning *in vivo* [44-46], it is clear that nucleosomes have certain sequence preference for their positioning. The region occupied by the centre of the nucleosome both *in vivo* and *in vitro* was shown to exhibit a significant increase in G/C usage, whereas A/T usage increases towards the nucleosome flanking regions [43]. Elements with such nucleotide composition were proposed to act as "container" sites able to produce a strongly positioned nucleosome [43], which then serves as a barrier for phasing of adjacent nucleosomes. On the other hand, a finer-scale 10 bp periodicity in A/T and G/C containing dinucleotides was found along the nucleosome-bound DNA and was proposed to contribute to precise positioning and/or rotational setting of DNA on nucleosomes [44,47].

How the nucleosome positioning pattern found around gene promoters is established and whether it requires active transcription by RNAPII machinery is still debated. There is evidence for both transcription-independent DNA sequence-driven [48], and transcriptional activity-aided nucleosome organisation [43], suggesting that there might not be a single mechanism responsible for nucleosome positioning at all promoters, but might be dependent on the type of the promoter itself.

4.2. DNA methylation and epigenetic features of CpG island promoters

In the scope of gene regulation, the term epigenetics refers to functionally relevant changes to the genome that influence gene expression without altering the underlying DNA sequence (genetic information). These can be chemical modifications to either DNA or histone proteins, which mediate both heritable changes in gene activity and long-term alterations in the transcriptional potential that are not necessarily heritable.

The best-studied epigenetic modification acting directly on DNA is methylation of cytosine, which in vertebrates occurs mainly in the CpG dinucleotide context. DNA methylation is

essential for normal development and is involved in several key processes including X-chromosome inactivation, genomic imprinting and suppression of repetitive elements [49]. *De novo* methylation occurs mainly during embryonic development, but it can also happen in adult cells due to aging or carcinogenesis. Majority of CpG dinucleotides in vertebrate genomes are methylated, except those located within CGIs. A small proportion of CGIs become methylated during development causing permanent silencing of associated promoters and ensuring lineage-specific expression of developmental regulatory genes [50]. There are several mechanisms by which CpG methylation mediates gene silencing: 1) methylated cytosines can alter binding sites for transcriptional activators and exclude them from binding [51], 2) mCpG can serve as a marker for methyl-cytosine binding domain proteins, which recruit co-repressor protein complexes that induce chromatin compaction [52] and 3) methylation directly increases affinity of certain sequences for histone octamer, thus increasing nucleosome occupancy and stability at promoters [53].

4.3. Promoters are marked by specific histone modifications

Unlike DNA, histones are subject to hundreds of covalent modifications, including acetylation, methylation, phosphorylation, and ubiquitination. These occur primarily at specific positions within the amino-terminal histone "tails", which emanate from the nucleosome core. Among various modifications, lysine acetylation and methylation are the most studied and best understood. Lysine acetylation almost always correlates with chromatin accessibility and transcriptional activity, and histone H3 lysine 27 acetylation (H3K27ac) was shown to mark active promoters and distal regulatory elements [54,55]. Tri-methylation of histone H3 lysine 4 (H3K4me3) and H3 lysine 36 (H3K36me3) are both associated with transcribed chromatin; however, H3K4me3 marks promoter regions, whereas H3K36me3 is found along the body of transcribed genes [41,56]. Unlike promoters, which are tri-methylated at H3 lysine 4, enhancers were shown to be mono-methylated [57]. Although these histone modifications in general correlate with transcriptional activity, it has been recently shown that transcription can occur in the absence of these canonical marks of active chromatin in *Drosophila* and worm [58]. In contrast to these active marks, tri-methylation of H3 lysine 9 (H3K9me3), H3 lysine 27 (H3K27me3) and H4 lysine 20 (H4K20me3) generally correlate with repression. H3K9me3 and H4K20me3 are marks of constitutive heterochromatin, a tightly packed repressive form of chromatin at repetitive portions of chromosomes [56]. Broad domains of H3K27me3 mark loci of transcriptionally silent developmental regulator genes in embryonic stem cells (ESC) [59]. The same loci were shown to contain punctuated H3K4me3 marks localised at promoters even though they were not

transcribed [59,60], suggesting that these "bivalent" domains silence developmental genes in ESCs while keeping them poised for activation.

Even from the very limited set of modifications described above, it is evident that the possibilities of marking genomic loci with various histone modifications and their combinations are enormous. It was proposed that specific combinations of modifications at given locus form a so called "histone code", which is read by other proteins to bring about distinct downstream events [61]. High-resolution mapping of numerous histone modifications in multiple cell types contributed to detection of most common combinations and associated functional genomic elements [62-64] and allowed segmentation of the genome into distinct domains based on the levels of various modifications [62,63,65]. Although specific histone modification combinations generally reflect the identity of the underlying DNA element, recent study has shown that actual levels of modification do not necessarily reflect the predicted regulatory activity [66].

5. Architecture and functional specialisation of Metazoan promoters

5.1. Core promoter elements and TSS selection

The "textbook" model of an RNAPII promoter has an A/T-rich DNA sequence (the TATA-box) approximately 30 bp upstream of the TSS, which in turn overlaps an initiator sequence (Inr) (Fig. 1). Assembly of a PIC at such promoters is initiated by TFIID binding to the TATA-box, Inr sequence and/or other sites [2]. TFIID is a multi-protein complex comprising the TATA-box binding protein (TBP) and more than 10 distinct TBP-associated factors (TAFs) [1]. TBP is a crucial component that recognises and binds the TATA-box motif [67], initiating subsequent PIC assembly and RNAPII recruitment. Once the PIC has assembled, the region around the TSS melts to provide a template strand for RNAPII, which occurs 25–30 bp downstream of the TATA-box in all eukaryotic model organisms studied so far, except in budding yeast, where this distance can vary [68]. Where present, the TATA-box seems to be the sole determinant of the TSS position, and initiation will occur at the distance set by the TATA-box regardless of the sequence around the site of initiation.

Although the TATA-box is a well-known core promoter motif it is present only in the minority (<15%) of mammalian promoters [12,69]. A more abundant, yet also not universal, metazoan

core promoter element is the initiator (Inr), which directly overlaps the TSS [70]. The consensus sequences of *Drosophila* and vertebrate Inr differ to some extent (Fig. 1), however in both cases they are bound by the homologous TAFs within the TFIID complex, which include TAF1 and TAF2 [2]. The common characteristic of the Inr element is the pyrimidine (C or T) / purine (A or G) motif (i.e. YR) positioned -1/+1 bp relative to the TSS, so that the purine is the first transcribed nucleotide [2,12]. Inr element often occurs in combination with either TATA-box [71], or with another core promoter element located downstream of the TSS, the downstream promoter element (DPE) [72]. They act synergistically to increase the efficiency of transcription by providing additional recognition sites for TFIID components and allowing cooperative TFIID binding.

The DPE was discovered in the analysis of TATA-less promoters in *Drosophila* [72] and was suggested to be conserved in humans [73]. This element acts in conjunction with the Inr, and the core sequence of the DPE is located at precisely +28 to +32 bp relative to the +1 nucleotide in the Inr motif [74]. This strict requirement for Inr–DPE spacing is essential for cooperative binding of TFIID, thus DPE and Inr function together as a single core promoter unit. Transcription initiation from DPE-containing promoters is dependent on TAFs, specifically TAF6 and TAF9, which were shown to bind DPE [1].

The TFIIB recognition element (BRE) is the only well-characterized core promoter motif bound by a factor other than TFIID. It was initially identified as a sequence immediately upstream of a subset of TATA-box elements [75]; however, an additional TFIIB recognition site, the downstream BRE, was found immediately downstream of the TATA box [76]. Several studies have shown that TFIIB plays a central role in transcription start site selection in both yeast and human [77]. Multiple mutations in TFIIB were found to cause a shift in the TSS selection, suggesting its role in precise positioning of RNAPII catalytic site at some core promoters [78]. BRE elements often occur in conjunction with the TATA-box and the observed spacing between TATA-box and TSS is a result of interaction between TBP, TFIIB and RNAPII, where TFIIB plays a central role in determining the spacing.

Despite the prevalence of CpG island-associated promoters, the precise mechanisms of their core promoter function are not well understood. One common feature of CGIs is the presence of multiple binding sites for transcription factor Sp1 [79]. Sp1 contributes to the maintenance of the hypomethylated state of CGIs and may work in concert with the general transcription machinery to support nucleation of the PIC [79]. TSSs are often located 40–80 bp downstream of the Sp1 sites, which suggests that Sp1 may direct the basal machinery to form PIC within a loosely defined

downstream window [80]. One possibility is that TFIID subunits that are capable of core promoter recognition then interact with the sequences within that window that are most compatible with their DNA recognition motifs, such as Inr element, to specify the exact TSS.

Initial studies suggested that the basal transcription machinery is largely invariant across different cell types and conditions. However, an increasing number of tissue-specific isoforms of TAFs as well as additional members of the TBP protein family such as TBP-related factors (TRFs) have been identified in Metazoa and found to form distinct TFIID-related complexes that can function at distinct core promoters [81,82]. Interestingly, many of these factors are involved in germ cell development [83]. The variability in basal transcription machinery composition might require different mechanisms for core promoter recognition leading to distinct patterns of TSS selection.

5.2. Nucleosome positioning and epigenetic features of promoter architectures

Distinct chromatin structure and histone modifications have been associated with active promoters. Both in yeast and Metazoa, the region immediately upstream of the TSS is marked as DNase I hypersensitive site, suggesting that it is a region of open chromatin depleted of nucleosomes [84]. This nucleosome-free region makes core promoter elements more accessible and facilitates PIC assembly and RNAPII recruitment. The accessibility of the promoter was shown to correlate with mRNA abundance [84].

The NFR is flanked by two nucleosomes, the first upstream or -1 nucleosome and the first downstream or +1 nucleosome, whose positioning can be more or less precise depending on the type of the promoter [34,85]. How the transcription initiation machinery contends with the +1 nucleosome seems to be different across different types of promoters. Precise mapping of PIC components in yeast showed that TFIID–TAF complex engages and is positioned by the +1 nucleosome at TATA-less promoters, whereas TATA-box containing promoters are largely depleted of TAFs and mediate PIC positioning through TBP and TFIIB interactions with the DNA [68]. Thus, in TATA-box promoters the +1 nucleosome can often overlap the TSS. Similarly, it was shown that at many promoters in *Drosophila* the +1 nucleosome resides >50 bp downstream of the TSS, where it engages with the paused RNAPII [42], further suggesting active role of the +1 nucleosome in transcriptional machinery positioning and RNAPII pausing.

Another important feature of nucleosomes flanking the TSS is the presence of specific histone variants. The H2A.Z variant was shown to be associated with promoters in both yeast and metazoa [41,42]; however, in yeast both -1 and +1 nucleosomes incorporate H2A.Z , whereas in *Drosophila* this variant is found exclusively in the +1 and additional downstream nucleosomes [42].

Histone variant H3.3 was also found to be enriched at promoters, where it was present almost exclusively together with H2A.Z. These H3.3/H2A.Z double variant–containing nucleosomes mark promoters and other regulatory regions and are surprisingly found within NFRs [86] which should by definition be devoid of nucleosomes. However, it seems that they are very unstable and thus not detected under the conditions normally used in nucleosome preparations [86]. This instability might facilitate the access of transcription factors to promoters and other regulatory sites *in vivo*.

Promoter-associated nucleosomes are also subject to various histone modifications that were shown to correlate with promoter activity [41,56,62,63]. The best-studied modifications associated with active promoters are H3K4me3 and H3K27ac, where H3K27ac level seems to be positively correlated with the level of expression, whereas H3K4me3 can be present on promoters that are not actively transcribing, but are poised for activation [59,62,63]. It was shown that basal transcription factor TFIID directly binds to the H3K4me3 mark via specific domain of TAF3 [87], which suggests that H3K4me3 might play an important role in defining core promoter. TAF3-mediated binding of TFIID to H3K4me3-marked nucleosomes could serve either to anchor TFIID to already activated promoters or to recruit TFIID during promoter activation. Interestingly, TAF3-H3K4me3 interaction seems to be more important for activation of TATA-less promoters, implying the importance of this mechanism for activation of promoters lacking canonical core promoter DNA elements [87]. However, it has been recently shown that transcription can occur in the absence of H3K4me3 and H3K27ac in *Drosophila* and worm [58], and this seems to be a distinctive feature of temporally regulated developmental genes, separating them from ubiquitously transcribed genes, which in contrast show high levels of these histone modifications.

Because many PIC components, including TFIID, have nucleosome-binding subunits, positioned nucleosomes might define the location of the TSS by positioning the PIC. The conventional view is that most genes contain a predominant TSS, the location of which is defined by core promoter elements [88]. However, many promoters lack any of the known core promoter elements and the question remains how the transcription machinery establishes the location of the TSS at those promoters. A model has been proposed in which TFIID complex binds to methylated (and acetylated) nucleosomes and recruits TBP to promoters [89]. TBP in turn binds TFIIB and places it immediately downstream towards the TSS. Since TFIIB was shown to dictate TSS selection [77], this model would explain how TSS positioning could be directed in part by TFIID bound to nucleosomes.

5.3. Promoter classes and modes of regulation

Early studies on individual promoters that led to the discovery of various core promoter elements already suggested substantial promoter heterogeneity. Some combinations of core promoter elements were observed more often than others, defining different structural and functional types of promoters. For instance, the TATA-box and DPE are rarely found together, but each of them is often associated with an Inr element [5,72,74]. Furthermore, the TATA-box containing and the DPE containing promoters appear to be functionally different, responding to distinct distal regulatory elements [90].

Genome-wide mapping of promoters and promoter-associated features allowed comprehensive analysis of promoter structure and function and their classification based on underlying sequence, chromatin, transcription initiation and expression specificity characteristics. The underlying sequence composition analysis revealed that mammalian promoters segregate naturally into two classes by CpG dinucleotide content: high-CpG and low-CpG promoters [9]. The former class is characterised by the overlap with CpG islands, thus they are also referred to as CGI-associated promoters. High resolution mapping of TSSs by CAGE distinguished two major classes of promoters based on the TSS distribution [12]. "Sharp" or "focused" promoters have a single well-defined TSS and are often associated with a TATA-box precisely positioned ~30 bp upstream of the TSS [12,24]. These classical "textbook" promoters represent only a minority of mammalian promoters and are commonly associated with tissue-specific genes and high conservation across species. Many TFs show distinct spatial biases with respect to TSS location and seem to be important contributors to the accurate prediction of single-peak TSSs [91]. The majority of mammalian promoters, however, comprise a second class of "broad" or "dispersed" promoters, characterised by multiple equally used TSSs distributed across a broader region [12], challenging the traditional definition of a gene and its precisely defined TSS. This class is strongly associated with CpG islands and ubiquitously expressed genes, however promoters of key developmental regulators were also found to belong to this class [92].

High resolution TSS mapping by PET [22] and CAGE [25] in *Drosophila* revealed analogous transcription initiation patterns, separating promoters into "sharp" and "broad" class. Unlike mammalian genome, the fly genome does not contain CpG islands; however, the two promoter classes were shown to be associated with distinct core promoter elements. The positionally restricted canonical core promoter elements, including TATA-box, Inr, DPE and MTE, were specifically enriched in sharp promoters [22,93]. When comparing across other *Drosophila* genomes, elements in broad promoters had lower levels of conservation than those in sharp

promoters [93]. Furthermore, the distinct promoter classes in fly were associated with the same functional categories of genes and showed similar expression specificity patterns as in mammals [12,22,93], suggesting functional conservation of the observed promoter classes across Metazoa. Interestingly, the distinct promoter classes were recently shown to respond to regulation by different sets of distal-acting enhancers, separating the housekeeping and developmental transcriptional programs in *Drosophila* [14] and emphasizing the importance of core promoters in transcriptional regulation during development.

Genome-wide analyses of various promoter-associated features provided further insight into structural and functional differences between CGI and non-CGI promoters in mammals. In pluripotent ES cells, vast majority of CGI promoters are associated with H3K4me3 enrichment [56], suggesting that they are targets of trithorax-group proteins, which catalyse the deposition of this mark. These promoters have a potential to drive transcription, unless they are actively repressed by Polycomb group proteins (PcG), which deposits repressive H3K27me3 mark and creates bivalent domains at key developmental genes and poises them for activation [59]. The ones that are not repressed tend to be ubiquitously expressed. In contrast, CpG-poor promoters seem to be inactive by default, independent of repression by PcG proteins, and may instead be selectively activated by cell-type- or tissue-specific factors [56]. This is further corroborated by the observation that CpG promoters are associated with RNAPII across multiple cell types, whereas non-CpG promoters acquire active chromatin marks and RNAPII binding in a tissue-dependent way [94]. The two promoter classes also differ in the nucleosome occupancy and the requirement for nucleosome remodelling complexes for their activation upon various external stimuli [95]. Taken together, this strongly suggests that CpG and non-CpG promoters in mammals are subject to distinct modes of regulation.

Unlike CGI and non-CGI promoter classification, which is vertebrate-specific, the corresponding sharp and broad promoter classes defined based on transcription initiation patterns are conserved across Metazoa [12,22,93]. These promoter classes are significantly differentiated by nucleosome organization and chromatin structure in both fly and mammals. Broad promoters display closer association with well-positioned nucleosomes and activating histone marks downstream of the TSS and have a more clearly defined NFR upstream, while sharp promoters have a less organized nucleosome structure and higher RNAPII presence [85].

Based on the configuration of promoter signals, TSS patterns, nucleosome positions and their epigenetic marks, and function of the associated gene, a unifying classification of Metazoan promoters into three main classes was proposed [34] (Fig. 3). Type I promoters are most often

used for genes that are specifically expressed in terminally differentiated peripheral tissues of an adult. They are characterised by a sharp transcription initiation pattern and are often associated with a TATA-box or other core promoter elements positionally restricted to the well-defined TSS in both mammals and fly. In mammals they are characterised by low CpG content and tend to have key regulatory inputs close to their promoters [96]. On chromatin level, Type I promoters are characterised by less-ordered nucleosomes [85], which can often cover the TSS; with H3K4me3 generally present downstream of the TSS when they are active and no RNAPII binding when they are not active [94]. However, a recent study suggests that these promoters might be active in the absence of canonically active histone modifications and proposes that for such promoters regulation by transcription factors has a more important regulatory role than chromatin marks [58]. Type II promoters are associated with ubiquitously active "housekeeping" genes and have broad promoter architecture with multiple TSSs spread across a wide region. In mammals, they tend to have a single CpG island covering the transcription initiation region, whereas in *Drosophila* they are associated with a distinct set of weaker core promoter elements [97]. The TSSs are located within a NFR and are flanked by two well-positioned nucleosomes that harbour active histone marks in all cell and tissue types, which seems to be associated with the stable production of RNA [58]. Type III promoters are characteristic of genes with expression that is developmentally regulated and coordinated across multiple cells. They share several characteristics with type II promoters, including a broad transcription initiation pattern and a well-defined NFR with positioned flanking nucleosomes, but also exhibit systematic differences that set them apart from the ubiquitously expressed class. The width of their transcription start region tends to be even broader than in Type II promoters [25]. Although their association with CpG islands in mammals is similar to type II promoters, developmental genes have longer or multiple CpG islands that often extend into the gene body [92]. The most prominent differences between type III and type II promoters are observed at the chromatin level. Developmental genes have a number of features that are associated with repression by PcG proteins, including wide distribution of PcG protein binding and both H3K27me3 and H3K4me3 marks, which create bivalent domains in ESCs [59]. Type III promoters are responsive to long-range regulation and can receive and integrate regulatory input from distal enhancers. They are often surrounded by arrays of highly conserved non-coding elements (HCNEs) that act as enhancers ensuring precise spatial and temporal expression of those key developmental regulators [92].

Studies in *Drosophila* and mammals have suggested that protein-coding genes with ubiquitous high expression whose protein products are components of translation machinery (ribosomal proteins, translation elongation and initiation factors) might have a separate, fourth architecture,

characterised by a pyrimidine-rich TCT initiator. This initiator motif is common to this functional category of genes in both *Drosophila* and mammalian genomes. It is characterised by a "sharp" transcription initiation pattern and, in *Drosophila*, it does not seem to contain a TATA box or any other known fixed-spacing motifs [6]. It has recently been reported that the PIC at this type of promoters lacks TBP, whose place is taken by the structurally related TRF2 [82]. Mammalian TCT promoters occasionally contain a canonical TATA box, but at present it cannot be excluded that this is due too multiple overlapping promoter architectures that are used independently (see below).

5.4. Promoter usage dynamics across cell types and developmental stages

The traditional view of a gene with its precisely defined and fixed TSS has been first challenged by the findings that many genes can be transcribed from multiple promoters (alternative promoters) producing functionally diverse transcripts [98,99]. Differential utilization of alternative promoters plays a critical role in regulating gene expression in a spatial, temporal or lineage-specific manner. This can be achieved by use of a distinct combination of core promoter elements in the alternative promoters [100,101]. Moreover, studying 5' ends of individual mRNAs genome-wide by CAGE, revealed that the transcription can start at multiple closely spaced TSSs within a single promoter [12], further increasing the diversity of produced transcripts. The closely spaced individual start sites can be associated with different core promoter elements and their activation can be dependent on distinct GTFs [102].

The complexity of transcription initiation in eukaryotic genomes is also seen in the bidirectional promoter arrangements, which in human genome comprise more than 10% of promoters [103]. Bidirectional promoters are associated with broad transcription start regions overlapping a CGI and display a mirror sequence composition [104]. The transcription from bidirectional promoters can be differentially regulated in the two directions [68], suggesting that the promoter elements and features can overlap in the same locus and be differentially interpreted by the RNAPII complexes transcribing independently in the opposite directions. Thus, bidirectional promoters are a good example of overlapping transcription initiation codes, which are differentially interpreted in different regulatory contexts.

Differential utilisation of promoter types has been observed across various contexts. For instance, in *Drosophila* embryonic development promoters of maternally inherited transcripts showed differences in motif composition compared to zygotically active promoters [93]. In addition, many genes with maternally inherited transcripts were found to have alternative promoters utilized later

in the development [93]. High-resolution quantitative mapping of TSSs across multiple human and mouse tissue types revealed substantial dynamics even at the level of individual TSSs within the same core promoter [105]. TSS selection within many CGI-associated broad promoters varies among tissues producing positional or regional bias in promoter usage [105]. This fine-scale regulation of transcription initiation events at the single base-pair level is likely related to epigenetic transcriptional regulation.

5.5. Overlapping transcription initiation codes: thousands of 2-in-1 promoters in vertebrate genomes

Mapping of precise TSSs across numerous mouse and human cell types by the FANTOM consortium provided evidence that RNA polymerase II has slightly different preference for TSS selection in different contexts, manifested as different positional distribution of transcription initiation events in "broad" promoters [27,105]. However, no clear rules or "codes" for TSS selection were evident from these analyses. By systematically analysing transcription initiation patterns and underlying sequence features in early development of zebrafish, a recent study revealed that the transcription initiation "code" in transcribing oocyte is different from that in somatic cells of the developing embryo, with different sequence elements that guide TSS selection at the promoter [26]. Most remarkably, the study showed that thousands of promoters that are active in both oocyte and somatic cells – including the "housekeeping" promoters – have both sets of promoter determinants, most often intertwined on the same physical stretch of DNA (Fig. 4). The oocyte-specific TSS selection is motif dependent, and the transcription always starts at a fixed distance from a weak TATA-like element (W-box). While sharing main features with initiation from a canonical TATA-dependent promoter, the oocyte-dependent promoters can have multiple W-boxes, each with its TSS ~30bp downstream of it, resulting in a composite sharp promoter architecture that gives an appearance of a broad promoter. On the other hand, the somatic TSS selection from the same promoters in the developing embryo is related to the stable position of first downstream (+1) nucleosome that determines the "catchment area" within which transcription can start at multiple TSSs resulting in a broad promoter architecture. Nonetheless, the transcription is still preferably initiated at YR dinucleotides at [-1,+1] positions, corresponding to loose vertebrate Initiator consensus sequence [12], and the one at optimal distance from the +1 nucleosome is used most frequently (i.e. it is the "dominant peak" of a broad promoter). Remarkably, the position of the dominant TSS alone is highly predictive of the +1 nucleosome position and reveals the presence of a sequence pattern characteristic for nucleosome bound

DNA downstream of the TSS in both zebrafish [26] and human [27], further corroborating the tight relationship between nucleosome positioning and TSS selection in broad promoters.

Further conclusions about how common are the alternative and overlapping transcription initiation "codes" in other cell types or organisms other than vertebrates are currently limited by the lack of the precise TSS data. There is evidence for oocyte-specific TSS code in another fish species (Haberle and Lenhard, unpublished), as well as a smaller-scale promoter code change during spermatogenesis in mouse [106]. In contrast, the global TSS use patterns seem remarkably stable across different somatic cell types, although differential TSS selection is evident at individual promoters between specific cell types [27]. Also, the purpose of a separate TSS selection code in oocyte is unclear at present: it may be used as an efficient way of genome-wide change of transcriptional repertoire between the oocyte and somatic cells – the most dramatic of such changes in the life cycle of Metazoa [107].

The overlapping transcription codes impose an additional layer of complexity to the genome wide computational analyses of promoters and their classification. New approaches are needed to detect and disentangle potential multiple and independent sets of promoter elements before attempting to classify them and understand their structure and function.

6. Diverse promoter architectures enable complex regulatory landscapes

Most of the regulatory content of a metazoan genome lies outside of proximal promoters [108] and tends to be contained within enhancers, which seem to be a predominant type of functional elements in the non-coding portion of the genome. They are characterised by clusters of binding sites for many different TFs and chromatin regulators [109,110]. Transcriptional activation by enhancers is temporally and spatially restricted and produces highly specific expression patterns during development [111].

Enhancers do not necessarily act on the closest promoter but can bypass neighbouring genes to regulate genes located more distantly along a chromosome, further increasing the complexity of the distal regulatory interactions within the genome. Given the nonlinear arrangement of developmental genes and their enhancers, a fundamental question is - how is the specificity between enhancer and its target promoter achieved? Several models have been proposed to

describe how enhancers may communicate with their target gene promoter [112]. Currently the most plausible model supported by both theoretical [113] and experimental [114,115] observations is the "looping" model in which the remote enhancer "loops out" the intervening DNA to reach the target promoter. It was shown that the formation of these chromatin loops depends on sequence-specific TFs bound to the enhancer and the promoter [115]. It appears that the enhancer loops form prior to gene activation and stably associate with paused RNAPII at promoters, keeping this loop topology ready for rapid activation of transcription by recruitment of additional factors [116]. The formation of chromatin loops brings the enhancer and its target promoter into close physical proximity in the nucleus and this feature is utilised by chromatin conformation capture experimental approaches [117] to detect long-range interactions genome-wide [118] and to identify target promoters of specific regulatory elements. However, the knowledge about the specificity of promoter-enhancer interactions is still very limited.

There is growing evidence that the features of the target promoter determine its responsiveness to distal regulatory elements within accessible chromosomal domain. For instance, it was shown that the presence of specific core promoter elements in *Drosophila* makes promoters responsive to distinct enhancers [119]. A recent functional study of enhancer activity genome-wide revealed specificity of enhancers towards either housekeeping or developmental core promoters that differ in their core promoter elements, separating two major transcriptional programs in *Drosophila* [14]. Furthermore, tightly regulated key developmental genes contained within large syntenic blocks in vertebrates [120] were shown to differ in various sequence, chromatin and transcriptional promoter features from neighbouring bystander genes, which likely specifies them as a target of regulation by surrounding HCNEs [92] (Fig. 5). These observations highlight the important functional role of the core promoter as an active participant in the long-range gene regulation.

7. Open questions and perspectives

This review gives an overview of the growing evidence that specific core promoter architectures play a central role in determining how a gene is regulated during development and differentiation. The architecture will determine whether the gene will be responsive to long-range regulatory inputs, where the majority of regulatory input is located with respect to the TSS [121], whether its transcriptional output will be stable or occur in bursts [122], and which epigenetic modifications will be present when the promoter is active or repressed. Architectural differences between promoters of different functional categories of genes appear to be ancient (for evidence of

different promoter types in yeast see e.g. [123] and [124]), and separate well constitutively active from regulated/inducible genes. Developmental promoters likely evolved from one of the ancestral classes by acquiring ability to integrate a large number of regulatory inputs in a manner easily malleable by selection forces. On the other hand, it is still not known how widespread is the utilisation of different transcription initiation codes discovered between oocyte and somatic cells that overlap on thousands of, mostly ubiquitously active, vertebrate promoters [26]. It will be interesting to find at what point in the evolution of Metazoa, or earlier, has this feature been acquired and what role it plays in distinguishing the global properties of transcription and its regulation between somatic cells and the germline.

Classification of promoter architectures and the characterisation of functionally equivalent architectures in distantly related species still remains to be done. Wider availability of the CAGE protocol and comparative promoterome analysis should enable the discovery of a finite number of promoter classes and serve as a starting point for their functional and mechanistic characterisation.

References

[1] M.C. Thomas, C.-M. Chiang, The general transcription machinery and general cofactors, Crit. Rev. Biochem. Mol. Biol. 41 (2006) 105–178. doi:10.1080/10409230600648736.

[2] S.T. Smale, J.T. Kadonaga, The RNA polymerase II core promoter, Annu. Rev. Biochem. 72 (2003) 449–479. doi:10.1146/annurev.biochem.72.121801.161520.

[3] J.T. Kadonaga, Perspectives on the RNA polymerase II core promoter, Wiley Interdiscip Rev Dev Biol. 1 (2012) 40–51. doi:10.1002/wdev.21.

[4] G.A. Maston, S.K. Evans, M.R. Green, Transcriptional Regulatory Elements in the Human Genome, Annu. Rev. Genom. Human Genet. 7 (2006) 29–59. doi:10.1146/annurev.genom.7.080505.115623.

[5] N.I. Gershenzon, I.P. Ioshikhes, Synergy of human Pol II core promoter elements revealed by statistical sequence analysis, Bioinformatics. 21 (2005) 1295–1300. doi:10.1093/bioinformatics/bti172.

[6] T.J. Parry, J.W.M. Theisen, J.-Y. Hsu, Y.-L. Wang, D.L. Corcoran, M. Eustice, et al., The TCT motif, a key component of an RNA polymerase II transcription system for the translational machinery, Genes & Development. 24 (2010) 2013–2018. doi:10.1101/gad.1951110.

[7] M. Gardiner-Garden, M. Frommer, CpG islands in vertebrate genomes, Journal of Molecular Biology. 196 (1987) 261–282.

[8] F. Antequera, A. Bird, Number of CpG islands and genes in human and mouse, Proc. Natl. Acad. Sci. U.S.A. 90 (1993) 11995–11999.

[9] S. Saxonov, P. Berg, D.L. Brutlag, A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters, Proc. Natl. Acad. Sci. U.S.A. 103 (2006) 1412–1417. doi:10.1073/pnas.0510310103.

[10] H.K. Long, D. Sims, A. Heger, N.P. Blackledge, C. Kutter, M.L. Wright, et al., Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates, eLife. 2 (2013). doi:10.7554/eLife.00348.016.
[11] D. Takai, P.A. Jones, Comprehensive analysis of CpG islands in human chromosomes 21 and 22, Proc. Natl. Acad. Sci. U.S.A. 99 (2002) 3740–3745. doi:10.1073/pnas.052410099.
[12] P. Carninci, A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, J. Ponjavic, et al., Genome-wide analysis of mammalian promoter architecture and evolution, Nat. Genet. 38 (2006) 626–635. doi:10.1038/ng1789.
[13] J.E.F. Butler, J.T. Kadonaga, The RNA polymerase II core promoter: a key component in the regulation of gene expression, Genes & Development. 16 (2002) 2583–2592. doi:10.1101/gad.1026202.
[14] M.A. Zabidi, C.D. Arnold, K. Schernhuber, M. Pagani, M. Rath, O. Frank, et al., Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation, Nature. 518 (2015) 556–559. doi:10.1038/nature13994.
[15] S. Lubliner, I. Regev, M. Lotan-Pompan, S. Edelheit, A. Weinberger, E. Segal, Core promoter sequence in yeast is a major determinant of expression level, Genome Res. 25 (2015) 1008–1017. doi:10.1101/gr.188193.114.
[16] T.H. Kim, L.O. Barrera, M. Zheng, C. Qu, M.A. Singer, T.A. Richmond, et al., A high-resolution map of active promoters in the human genome, Nature. 436 (2005) 876–880. doi:10.1038/nature03877.
[17] Z. Wang, M. Gerstein, M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics, Nat. Rev. Genet. 10 (2009) 57–63. doi:10.1038/nrg2484.
[18] Y. Okazaki, M. Furuno, T. Kasukawa, J. Adachi, H. Bono, S. Kondo, et al., Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs, Nature. 420 (2002) 563–573. doi:10.1038/nature01266.
[19] Y. Suzuki, T. Tsunoda, J. Sese, H. Taira, J. Mizushima-Sugano, H. Hata, et al., Identification and characterization of the potential promoter regions of 1031 kinds of human genes, Genome Res. 11 (2001) 677–684. doi:10.1101/gr.164001.
[20] N.D. Trinklein, S.J.F. Aldred, A.J. Saldanha, R.M. Myers, Identification and functional analysis of human transcriptional promoters, Genome Res. 13 (2003) 308–312. doi:10.1101/gr.794803.
[21] Z. Zhang, F.S. Dietrich, Mapping of transcription start sites in Saccharomyces cerevisiae using 5' SAGE, Nucleic Acids Res. 33 (2005) 2838–2851. doi:10.1093/nar/gki583.
[22] T. Ni, D.L. Corcoran, E.A. Rach, S. Song, E.P. Spana, Y. Gao, et al., A paired-end sequencing strategy to map the complex landscape of transcription initiation, Nat. Methods. 7 (2010) 521–527. doi:10.1038/nmeth.1464.
[23] T. Shiraki, S. Kondo, S. Katayama, K. Waki, T. Kasukawa, H. Kawaji, et al., Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage, Proc. Natl. Acad. Sci. U.S.A. 100 (2003) 15776–15781. doi:10.1073/pnas.2136655100.
[24] J. Ponjavic, B. Lenhard, C. Kai, J. Kawai, P. Carninci, Y. Hayashizaki, et al., Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters, Genome Biol. 7 (2006) R78. doi:10.1186/gb-2006-7-8-R78.
[25] R.A. Hoskins, J.M. Landolin, J.B. Brown, J.E. Sandler, H. Takahashi, T. Lassmann, et al., Genome-wide analysis of promoter architecture in Drosophila melanogaster, Genome Res. 21 (2011) 182–192. doi:10.1101/gr.112466.110.
[26] V. Haberle, N. Li, Y. Hadzhiev, C. Plessy, C. Previti, C. Nepal, et al., Two independent transcription initiation codes overlap on vertebrate core promoters, Nature. 507 (2014) 381–385. doi:10.1038/nature12974.
[27] The FANTOM Consortium and the RIKEN PMI and CLST (DGT), A promoter-level mammalian expression atlas, Nature. 507 (2014) 462–470. doi:10.1038/nature13182.

[28] H. Takahashi, T. Lassmann, M. Murata, P. Carninci, 5' end-centered expression profiling using cap-analysis gene expression and next-generation sequencing, Nature Protocols. 7 (2012) 542–561. doi:10.1038/nprot.2012.005.
[29] M. de Hoon, Y. Hayashizaki, Deep cap analysis gene expression (CAGE): genome-wide identification of promoters, quantification of their expression, and network inference, BioTechniques. 44 (2008) 627–632. doi:10.2144/000112802.
[30] FANTOM Consortium, H. Suzuki, A.R.R. Forrest, E. van Nimwegen, C.O. Daub, P.J. Balwierz, et al., The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line, Nat. Genet. 41 (2009) 553–562. doi:10.1038/ng.375.
[31] E. Valen, G. Pascarella, A. Chalk, N. Maeda, M. Kojima, C. Kawazu, et al., Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE, Genome Res. 19 (2009) 255–265. doi:10.1101/gr.084541.108.
[32] P. Carninci, T. Kasukawa, S. Katayama, J. Gough, M.C. Frith, N. Maeda, et al., The Transcriptional Landscape of the Mammalian Genome, Science. 309 (2005) 1559–1563. doi:10.1126/science.1112014.
[33] G.J. Faulkner, Y. Kimura, C.O. Daub, S. Wani, C. Plessy, K.M. Irvine, et al., The regulated retrotransposon transcriptome of mammalian cells, Nat. Genet. 41 (2009) 563–571. doi:10.1038/ng.368.
[34] B. Lenhard, A. Sandelin, P. Carninci, Metazoan promoters: emerging characteristics and insights into transcriptional regulation, Nat. Rev. Genet. 13 (2012) 233–245. doi:10.1038/nrg3163.
[35] S. Djebali, C.A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, et al., Landscape of transcription in human cells, Nature. 488 (2012) 101–108. doi:10.1038/nature11233.
[36] R. Andersson, C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt, M. Boyd, et al., An atlas of active enhancers across human cell types and tissues, Nature. 507 (2014) 455–461. doi:10.1038/nature12787.
[37] Affymetrix/Cold Spring Harbor Laboratory ENCODE Transcriptome Project, Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs, Nature. 457 (2009) 1028–1032. doi:10.1038/nature07759.
[38] C. Nepal, Y. Hadzhiev, C. Previti, V. Haberle, N. Li, H. Takahashi, et al., Dynamic regulation of the transcription initiation landscape at single nucleotide resolution during vertebrate embryogenesis, Genome Res. 23 (2013) 1938–1950. doi:10.1101/gr.153692.112.
[39] L.J. Core, A.L. Martins, C.G. Danko, C.T. Waters, A. Siepel, J.T. Lis, Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers, Nat. Genet. 46 (2014) 1311–1320. doi:10.1038/ng.3142.
[40] G.-C. Yuan, Y.-J. Liu, M.F. Dion, M.D. Slack, L.F. Wu, S.J. Altschuler, et al., Genome-scale identification of nucleosome positions in S. cerevisiae, Science. 309 (2005) 626–630. doi:10.1126/science.1112178.
[41] A. Barski, S. Cuddapah, K. Cui, T.-Y. Roh, D.E. Schones, Z. Wang, et al., High-Resolution Profiling of Histone Methylations in the Human Genome, Cell. 129 (2007) 823–837. doi:10.1016/j.cell.2007.05.009.
[42] T.N. Mavrich, C. Jiang, I.P. Ioshikhes, X. Li, B.J. Venters, S.J. Zanton, et al., Nucleosome organization in the Drosophila genome, Nature. 453 (2008) 358–362. doi:10.1038/nature06929.
[43] A. Valouev, S.M. Johnson, S.D. Boyd, C.L. Smith, A.Z. Fire, A. Sidow, Determinants of nucleosome organization in primary human cells, Nature. 474 (2011) 516–520. doi:10.1038/nature10002.
[44] N. Kaplan, I.K. Moore, Y. Fondufe-Mittendorf, A.J. Gossett, D. Tillo, Y. Field, et al., The DNA-encoded nucleosome organization of a eukaryotic genome, Nature. 458 (2009) 362–

366. doi:10.1038/nature07667.
[45] Y. Zhang, Z. Moqtaderi, B.P. Rattner, G. Euskirchen, M. Snyder, J.T. Kadonaga, et al., Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo, Nat Struct Mol Biol. 16 (2009) 847–852. doi:10.1038/nsmb.1636.
[46] N. Kaplan, I. Moore, Y. Fondufe-Mittendorf, A.J. Gossett, D. Tillo, Y. Field, et al., Nucleosome sequence preferences influence in vivo nucleosome organization, Nat Struct Mol Biol. 17 (2010) 918–920. doi:10.1038/nsmb0810-918.
[47] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thåström, Y. Field, I.K. Moore, et al., A genomic code for nucleosome positioning, Nature. 442 (2006) 772–778. doi:10.1038/nature04979.
[48] R. Fenouil, P. Cauchy, F. Koch, N. Descostes, J.Z. Cabeza, C. Innocenti, et al., CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters, Genome Res. 22 (2012) 2399–2408. doi:10.1101/gr.138776.112.
[49] A. Bird, DNA methylation patterns and epigenetic memory, Genes & Development. 16 (2002) 6–21. doi:10.1101/gad.947102.
[50] W. Xie, M.D. Schultz, R. Lister, Z. Hou, N. Rajagopal, P. Ray, et al., Epigenomic Analysis of Multilineage Differentiation of Human Embryonic Stem Cells, Cell. 153 (2013) 1134–1148. doi:10.1016/j.cell.2013.04.022.
[51] S. Hu, J. Wan, Y. Su, Q. Song, Y. Zeng, H.N. Nguyen, et al., DNA methylation presents distinct binding sites for human transcription factors, eLife. 2 (2013) e00726. doi:10.7554/eLife.00726.
[52] L. Lande-Diner, J. Zhang, I. Ben-Porath, N. Amariglio, I. Keshet, M. Hecht, et al., Role of DNA methylation in stable gene repression, J. Biol. Chem. 282 (2007) 12194–12200. doi:10.1074/jbc.M607838200.
[53] C.K. Collings, P.J. Waddell, J.N. Anderson, Effects of DNA methylation on nucleosome stability, Nucleic Acids Res. 41 (2013) 2918–2931. doi:10.1093/nar/gks893.
[54] N.D. Heintzman, G.C. Hon, R.D. Hawkins, P. Kheradpour, A. Stark, L.F. Harp, et al., Histone modifications at human enhancers reflect global cell-type-specific gene expression, Nature. 459 (2009) 108–112. doi:10.1038/nature07829.
[55] M.P. Creyghton, A.W. Cheng, G.G. Welstead, T. Kooistra, B.W. Carey, E.J. Steine, et al., Histone H3K27ac separates active from poised enhancers and predicts developmental state, Proceedings of the National Academy of Sciences. 107 (2010) 21931–21936. doi:10.1073/pnas.1016071107.
[56] T.S. Mikkelsen, M. Ku, D.B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, et al., Genome-wide maps of chromatin state in pluripotent and lineage-committed cells, Nature. 448 (2007) 553–560. doi:10.1038/nature06008.
[57] N.D. Heintzman, R.K. Stuart, G. Hon, Y. Fu, C.W. Ching, R.D. Hawkins, et al., Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome, Nat. Genet. 39 (2007) 311–318. doi:10.1038/ng1966.
[58] S. Pérez-Lluch, E. Blanco, H. Tilgner, J. Curado, M. Ruiz-Romero, M. Corominas, et al., Absence of canonical marks of active chromatin in developmentally regulated genes, Nat. Genet. 47 (2015) 1158–1167. doi:10.1038/ng.3381.
[59] B.E. Bernstein, T.S. Mikkelsen, X. Xie, M. Kamal, D.J. Huebert, J. Cuff, et al., A bivalent chromatin structure marks key developmental genes in embryonic stem cells, Cell. 125 (2006) 315–326. doi:10.1016/j.cell.2006.02.041.
[60] M.G. Guenther, S.S. Levine, L.A. Boyer, R. Jaenisch, R.A. Young, A chromatin landmark and transcription initiation at most promoters in human cells, Cell. 130 (2007) 77–88. doi:10.1016/j.cell.2007.05.042.
[61] O.J. Rando, Combinatorial complexity in chromatin structure and function: revisiting the histone code, Current Opinion in Genetics & Development. 22 (2012) 148–155.

doi:10.1016/j.gde.2012.02.013.
[62] P.V. Kharchenko, A.A. Alekseyenko, Y.B. Schwartz, A. Minoda, N.C. Riddle, J. Ernst, et al., Comprehensive analysis of the chromatin landscape in Drosophila melanogaster, Nature. 471 (2011) 480–485. doi:10.1038/nature09725.
[63] J. Ernst, P. Kheradpour, T.S. Mikkelsen, N. Shoresh, L.D. Ward, C.B. Epstein, et al., Mapping and analysis of chromatin state dynamics in nine human cell types, Nature. 473 (2011) 43–49. doi:10.1038/nature09906.
[64] The ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome, Nature. 489 (2012) 57–74. doi:10.1038/nature11247.
[65] J. Ernst, M. Kellis, ChromHMM: automating chromatin-state discovery and characterization, Nat. Methods. 9 (2012) 215–216. doi:10.1038/nmeth.1906.
[66] J.C. Kwasnieski, C. Fiore, H.G. Chaudhari, B.A. Cohen, High-throughput functional testing of ENCODE segmentation predictions, Genome Res. (2014) gr.173518.114. doi:10.1101/gr.173518.114.
[67] S.K. Burley, The TATA box binding protein, Current Opinion in Structural Biology. 6 (1996) 69–75.
[68] H.S. Rhee, B.F. Pugh, Genome-wide structure and organization of eukaryotic pre-initiation complexes, Nature. 483 (2012) 295–301. doi:10.1038/nature10799.
[69] S.J. Cooper, N.D. Trinklein, E.D. Anton, L. Nguyen, R.M. Myers, Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome, Genome Res. 16 (2006) 1–10. doi:10.1101/gr.4222606.
[70] C. Yang, E. Bolotin, T. Jiang, F.M. Sladek, E. Martinez, Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters, Gene. 389 (2007) 52–65. doi:10.1016/j.gene.2006.09.029.
[71] S.T. Smale, D. Baltimore, The "initiator" as a transcription control element, Cell. 57 (1989) 103–113.
[72] T.W. Burke, J.T. Kadonaga, Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters, Genes & Development. 10 (1996) 711–724.
[73] T.W. Burke, J.T. Kadonaga, The downstream core promoter element, DPE, is conserved from Drosophila to humans and is recognized by TAFII60 of Drosophila, Genes & Development. 11 (1997) 3020–3031. doi:10.1101/gad.11.22.3020.
[74] A.K. Kutach, J.T. Kadonaga, The downstream promoter element DPE appears to be as widely used as the TATA box in Drosophila core promoters, Molecular and Cellular Biology. 20 (2000) 4754–4764.
[75] T. Lagrange, A.N. Kapanidis, H. Tang, D. Reinberg, R.H. Ebright, New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB, Genes & Development. 12 (1998) 34–44.
[76] W. Deng, S.G.E. Roberts, A core promoter element downstream of the TATA box that is recognized by TFIIB, Genes & Development. 19 (2005) 2418–2423. doi:10.1101/gad.342405.
[77] N.A. Hawkes, S.G. Roberts, The role of human TFIIB in transcription start site selection in vitro and in vivo, J. Biol. Chem. 274 (1999) 14337–14343.
[78] W. Deng, S.G.E. Roberts, TFIIB and the regulation of transcription by RNA polymerase II, Chromosoma. 116 (2007) 417–429. doi:10.1007/s00412-007-0113-9.
[79] M. Brandeis, D. Frank, I. Keshet, Z. Siegfried, M. Mendelsohn, A. Nemes, et al., Sp1 elements protect a CpG island from de novo methylation, Nature. 371 (1994) 435–438. doi:10.1038/371435a0.
[80] M.C. Blake, R.C. Jambou, A.G. Swick, J.W. Kahn, J.C. Azizkhan, Transcriptional initiation is controlled by upstream GC-box interactions in a TATAA-less promoter, Molecular and

Cellular Biology. 10 (1990) 6632–6641.
[81] F.J. Herrera, T. Yamaguchi, H. Roelink, R. Tjian, Core promoter factor TAF9B regulates neuronal gene expression, eLife. 3 (2014) e02559. doi:10.7554/eLife.02559.
[82] Y.-L. Wang, S.H.C. Duttke, K. Chen, J. Johnston, G.A. Kassavetis, J. Zeitlinger, et al., TRF2, but not TBP, mediates the transcription of ribosomal protein genes, Genes & Development. 28 (2014) 1550–1555. doi:10.1101/gad.245662.114.
[83] W. Akhtar, G.J.C. Veenstra, TBP2 is a substitute for TBP in Xenopus oocyte transcription, BMC Biol. 7 (2009) 45. doi:10.1186/1741-7007-7-45.
[84] J.R. Hesselberth, X. Chen, Z. Zhang, P.J. Sabo, R. Sandstrom, A.P. Reynolds, et al., Global mapping of protein-DNA interactions in vivo by digital genomic footprinting, Nat. Methods. 6 (2009) 283–289. doi:10.1038/nmeth.1313.
[85] E.A. Rach, D.R. Winter, A.M. Benjamin, D.L. Corcoran, T. Ni, J. Zhu, et al., Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level, PLoS Genet. 7 (2011) e1001274. doi:10.1371/journal.pgen.1001274.
[86] C. Jin, C. Zang, G. Wei, K. Cui, W. Peng, K. Zhao, et al., H3.3/H2A.Z double variant-containing nucleosomes mark "nucleosome-free regions" of active promoters and other regulatory regions, Nat. Genet. 41 (2009) 941–945. doi:10.1038/ng.409.
[87] M. Vermeulen, K.W. Mulder, S. Denissov, W.W.M.P. Pijnappel, F.M.A. van Schaik, R.A. Varier, et al., Selective Anchoring of TFIID to Nucleosomes by Trimethylation of Histone H3 Lysine 4, Cell. 131 (2007) 58–69. doi:10.1016/j.cell.2007.08.016.
[88] T. Juven-Gershon, J.T. Kadonaga, Regulation of gene expression via the core promoter and the basal transcriptional machinery, Developmental Biology. 339 (2010) 225–229. doi:10.1016/j.ydbio.2009.08.009.
[89] C. Jiang, B.F. Pugh, Nucleosome positioning and gene regulation: advances through genomics, Nat. Rev. Genet. 10 (2009) 161–172. doi:10.1038/nrg2522.
[90] J.E.F. Butler, J.T. Kadonaga, Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs, Genes & Development. 15 (2001) 2515–2519. doi:10.1101/gad.924301.
[91] M. Megraw, F. Pereira, S.T. Jensen, U. Ohler, A.G. Hatzigeorgiou, A transcription factor affinity-based code for mammalian transcription initiation, Genome Res. 19 (2009) 644–656. doi:10.1101/gr.085449.108.
[92] A. Akalin, D. Fredman, E. Arner, X. Dong, J.C. Bryne, H. Suzuki, et al., Transcriptional features of genomic regulatory blocks, Genome Biol. 10 (2009) R38. doi:10.1186/gb-2009-10-4-r38.
[93] E.A. Rach, H.-Y. Yuan, W.H. Majoros, P. Tomancak, U. Ohler, Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the Drosophila genome, Genome Biol. 10 (2009) R73. doi:10.1186/gb-2009-10-7-r73.
[94] L.O. Barrera, Z. Li, A.D. Smith, K.C. Arden, W.K. Cavenee, M.Q. Zhang, et al., Genome-wide mapping and analysis of active promoters in mouse embryonic stem cells and adult organs, Genome Res. 18 (2008) 46–59. doi:10.1101/gr.6654808.
[95] V.R. Ramirez-Carrozzi, D. Braas, D.M. Bhatt, C.S. Cheng, C. Hong, K.R. Doty, et al., A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling, Cell. 138 (2009) 114–128. doi:10.1016/j.cell.2009.04.020.
[96] H.G. Roider, B. Lenhard, A. Kanhere, S.A. Haas, M. Vingron, CpG-depleted promoters harbor tissue-specific transcription factor binding signals--implications for motif overrepresentation analyses, Nucleic Acids Res. 37 (2009) 6305–6315. doi:10.1093/nar/gkp682.
[97] U. Ohler, Identification of core promoter modules in Drosophila and their application in accurate transcription start site prediction, Nucleic Acids Res. 34 (2006) 5943–5950. doi:10.1093/nar/gkl608.
[98] T.A. Ayoubi, W.J. Van de Ven, Regulation of gene expression by alternative promoters,

The FASEB Journal. 10 (1996) 1–8.
[99] R.V. Davuluri, Y. Suzuki, S. Sugano, C. Plass, T.H.-M. Huang, The functional consequences of alternative promoter use in mammalian genomes, Trends Genet. 24 (2008) 167–177. doi:10.1016/j.tig.2008.01.008.
[100] S.K. Hansen, R. Tjian, TAFs and TFIIA mediate differential utilization of the tandem Adh promoters, Cell. 82 (1995) 565–575.
[101] B. Ren, T. Maniatis, Regulation of Drosophila Adh promoter switching by an initiator-targeted repression mechanism, Embo J. 17 (1998) 1076–1086. doi:10.1093/emboj/17.4.1076.
[102] G. Mizuguchi, C. Kanei-Ishii, T. Sawazaki, M. Horikoshi, R.G. Roeder, T. Yamamoto, et al., Independent control of transcription initiations from two sites by an initiator-like element and TATA box in the human c-erbB-2 promoter, FEBS Lett. 348 (1994) 80–88.
[103] N.D. Trinklein, An Abundance of Bidirectional Promoters in the Human Genome, Genome Res. 14 (2004) 62–66. doi:10.1101/gr.1982804.
[104] P.G. Engström, H. Suzuki, N. Ninomiya, A. Akalin, L. Sessa, G. Lavorgna, et al., Complex Loci in Human and Mouse Genomes, PLoS Genet. 2 (2006) e47. doi:10.1371/journal.pgen.
[105] H. Kawaji, M.C. Frith, S. Katayama, A. Sandelin, C. Kai, J. Kawai, et al., Dynamic usage of transcription start sites within core promoters, Genome Biol. 7 (2006) R118. doi:10.1186/gb-2006-7-12-r118.
[106] V. Haberle, A.R.R. Forrest, Y. Hayashizaki, P. Carninci, B. Lenhard, CAGEr: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses, Nucleic Acids Res. 43 (2015) e51–e51. doi:10.1093/nar/gkv054.
[107] W. Tadros, H.D. Lipshitz, The maternal-to-zygotic transition: a play in two acts, Development. 136 (2009) 3033–3042. doi:10.1242/dev.033183.
[108] R.E. Thurman, E. Rynes, R. Humbert, J. Vierstra, M.T. Maurano, E. Haugen, et al., The accessible chromatin landscape of the human genome, Nature. 489 (2012) 75–82. doi:10.1038/nature11232.
[109] L.A. Pennacchio, N. Ahituv, A.M. Moses, S. Prabhakar, M.A. Nobrega, M. Shoukry, et al., In vivo enhancer analysis of human conserved non-coding sequences, Nature. 444 (2006) 499–502. doi:10.1038/nature05295.
[110] L.A. Pennacchio, G.G. Loots, M.A. Nobrega, I. Ovcharenko, Predicting tissue-specific enhancers in the human genome, Genome Res. 17 (2007) 201–211. doi:10.1101/gr.5972507.
[111] A. Visel, J.A. Akiyama, M. Shoukry, V. Afzal, E.M. Rubin, L.A. Pennacchio, Functional autonomy of distant-acting human enhancers, Genomics. 93 (2009) 509–513. doi:10.1016/j.ygeno.2009.02.002.
[112] P. Kolovos, T.A. Knoch, F.G. Grosveld, P.R. Cook, A. Papantonis, Enhancers and silencers: an integrated and simple model for their function, Epigenetics Chromatin. 5 (2012) 1. doi:10.1186/1756-8935-5-1.
[113] S. Mukhopadhyay, P. Schedl, V.M. Studitsky, A.M. Sengupta, Theoretical analysis of the role of chromatin interactions in long-range action of enhancers and insulators, Proc. Natl. Acad. Sci. 19919-19924 (2011) 19919–19924. doi:10.1073/pnas.1103845108/-/DCSupplemental.
[114] B. Tolhuis, R.-J. Palstra, E. Splinter, F. Grosveld, W. de Laat, Looping and interaction between hypersensitive sites in the active beta-globin locus, Molecular Cell. 10 (2002) 1453–1465.
[115] I.K. Nolis, D.J. McKay, E. Mantouvalou, S. Lomvardas, M. Merika, D. Thanos, Transcription factors mediate long-range enhancer-promoter interactions, Proc. Natl. Acad. Sci. U.S.A. 106 (2009) 20222–20227. doi:10.1073/pnas.0902454106.
[116] Y. Ghavi-Helm, F.A. Klein, T. Pakozdi, L. Ciglar, D. Noordermeer, W. Huber, et al.,

Enhancer loops appear stable during development and are associated with paused polymerase, Nature. (2014) 1–22. doi:10.1038/nature13417.
[117] E. de Wit, W. de Laat, A decade of 3C technologies: insights into nuclear organization, Genes & Development. 26 (2012) 11–24. doi:10.1101/gad.179804.111.
[118] A. Sanyal, B.R. Lajoie, G. Jain, J. Dekker, The long-range interaction landscape of gene promoters, Nature. 489 (2012) 109–113. doi:10.1038/nature11279.
[119] P.G. Engstrom, S.J. Ho Sui, O. Drivenes, T.S. Becker, B. Lenhard, Genomic regulatory blocks underlie extensive microsynteny conservation in insects, Genome Res. 17 (2007) 1898–1908. doi:10.1101/gr.6669607.
[120] H. Kikuta, M. Laplante, P. Navratilova, A.Z. Komisarczuk, P.G. Engstrom, D. Fredman, et al., Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates, Genome Res. 17 (2007) 545–555. doi:10.1101/gr.6086307.
[121] E. Soler, C. Andrieu-Soler, E. de Boer, J.C. Bryne, S. Thongjuea, R. Stadhouders, et al., The genome-wide dynamics of the binding of Ldb1 complexes during erythroid differentiation, Genes & Development. 24 (2010) 277–289. doi:10.1101/gad.551810.
[122] D.M. Suter, N. Molina, D. Gatfield, K. Schneider, U. Schibler, F. Naef, Mammalian genes are transcribed with widely different bursting kinetics, Science. 332 (2011) 472–474. doi:10.1126/science.1198817.
[123] J.B. Zaugg, N.M. Luscombe, A genomic model of condition-specific nucleosome behavior explains transcriptional activity in yeast, Genome Res. 22 (2012) 84–94. doi:10.1101/gr.124099.111.
[124] V. Charoensawan, S.C. Janga, M.L. Bulyk, M.M. Babu, S.A. Teichmann, DNA sequence preferences of transcriptional activators correlate more strongly than repressors with nucleosomes, Molecular Cell. 47 (2012) 183–192. doi:10.1016/j.molcel.2012.06.028.
[125] U. Ohler, G.-C. Liao, H. Niemann, G.M. Rubin, Computational analysis of core promoters in the Drosophila genome, Genome Biol. 3 (2002) RESEARCH0087.

Figure Legends

Figure 1. Summary of most prevalent core promoter elements positionally constrained with respect to transcription start site (TSS; marked as +1 position). The location of elements relative to the TSS is shown as coloured boxes, where the colour indicates whether the element is *Drosophila*-specific (red), vertebrate-specific (blue) or common (purple). Associated sequence logos are based on motifs from [125] and [6] for *Drosophila* and motifs from the JASPAR database for vertebrates. The initiator motif (Inr) differs between *Drosophila* and vertebrates and both sequence logos are shown. Most promoters only have one or a few of these elements, and some elements are mostly found in certain species. BRE, TFIIB recognition element; DCE, downstream core element; DPE, downstream promoter element; Inr, initiator; MTE, motif ten element; TATA, TATA-box element; TCT, TCT initiator. IMPORTANT: hardly any real promoter contains all or even most of the above elements – on the contrary, different elements are associated with

different promoter architectures and their co-occurrence in individual promoters are strongly underrepresented compared to chance.

Figure 2. Schematic overview of Cap Analysis of Gene Expression (CAGE) method for identifying promoters at high resolution (redrawn based on [28]). By mapping exact 5' ends of complete cDNAs CAGE provides genome-wide single base-pair resolution map of transcription start sites and relative levels of transcripts initiated at each individual TSS.

Figure 3. Main classes of promoters as described in [34]. Metazoan promoters can broadly be divided into three groups based on their transcription initiation patterns (red), localisation of context-specific regulatory input (blue), sequence signals (gray) and nucleosome configuration around TSS (pink). Other less frequent promoter types associated with specific functional groups of genes have been described, such as TCT initiator containing promoters of translational machinery genes [6] (bottom).

Figure 4. Overlapping promoter "codes". Two independent promoter codes (shown in red and blue) overlap on thousands of promoters and guide differential TSS selection in the zebrafish oocyte and the somatic cells of the developing embryo [26]. The transcriptional machinery in the oocyte reads the "blue" code that consists of multiple A/T rich W-box motifs positioned ~30bp upstream of respective TSSs, resulting in composite sharp promoter architecture (top). In the embryo the "red" code is read instead, which restricts the TSS selection to a "catchment" area by the precisely positioned first downstream (+1) nucleosome aligned with a nucleosome positioning signal in the sequence. The transcriptional machinery initiates most frequently at the loose initiator motif (YR dinucleotide) at the optimal ~50bp position upstream of the +1 nucleosome (bottom).

Figure 5. The role of the core promoter type in long-range gene regulation. Tightly regulated key developmental genes (shown in red) contained within large syntenic blocks receive regulatory input from distal-acting enhancers. Promoters of these genes differ in various sequence, chromatin and transcriptional features from neighbouring bystander genes (shown in blue), which likely specifies them as a target of regulation by surrounding enhancers [92].

Figure 1

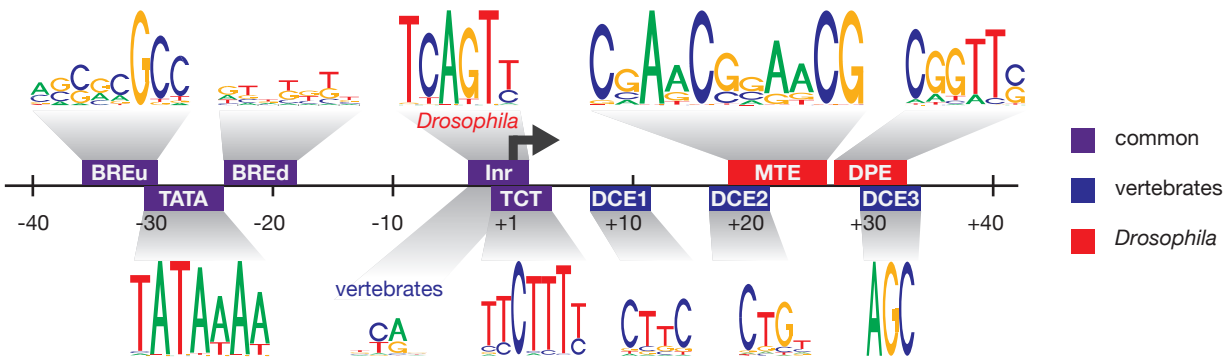

BREu
TATA
BREd
Inr
TCT
MTE
DPE
DCE1
DCE2
DCE3
-40
-30
-20
-10
+1
+10
+20
+30
+40
Drosophila
vertebrates
common
vertebrates
Drosophila

Figure 2

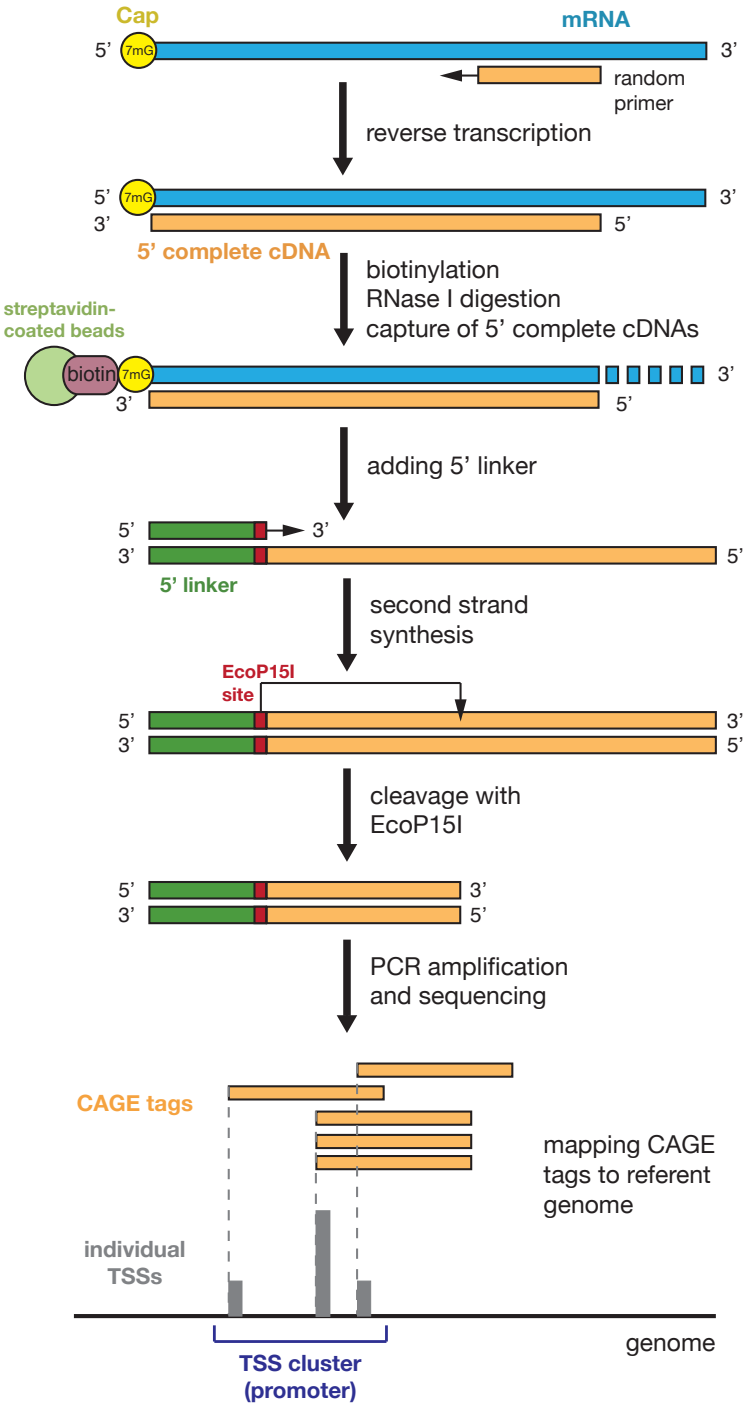
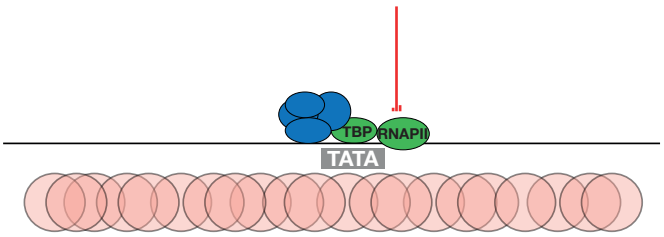

Cap
mRNA
5'
7mG
3'
random primer
reverse transcription
5' complete cDNA
biotinylation
RNase I digestion
capture of 5' complete cDNAs
streptavidin-coated beads
biotin
adding 5' linker
5' linker
second strand synthesis
EcoP15I site
cleavage with EcoP15I
PCR amplification and sequencing
CAGE tags
mapping CAGE tags to referent genome
individual TSSs
genome
TSS cluster (promoter)

Figure 3

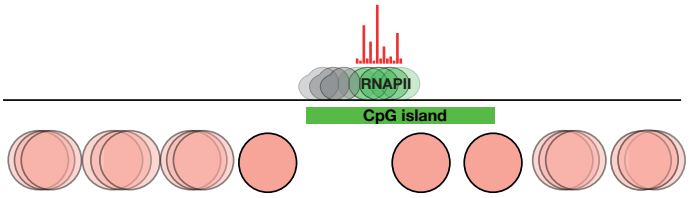
Adult tissue-specific

1) sharp transcription initiation pattern
2) context-specific regulatory input close to TSS
3) TATA-box ~30 bp upstream of TSS
4) disordered nucleosomes



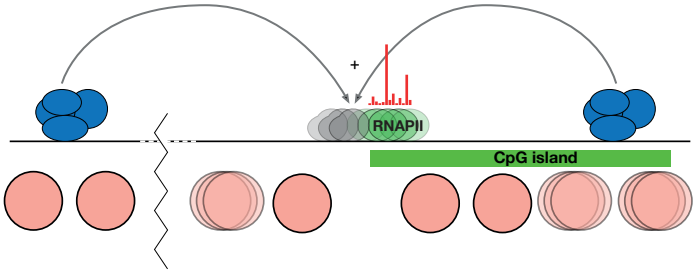
Ubiquitously expressed ("housekeeping")

1) broad transcription initiation pattern
2) no context-specific regulatory input
3) CpG island around TSS
4) nucleosome-free region with precisely positioned -1 and +1 nucleosomes



Developmentally regulated

1) broad transcription initiation pattern
2) context-specific regulatory input from distal enhancers
3) long CpG island(s) into gene body
4) nucleosome-free region with precisely positioned -1 and +1 nucleosomes



Translational machinery-specific

1) sharp transcription initiation pattern
2) no context-specific regulatory input
3) TCT initiator and CpG island
4) nucleosome-free region with precisely positioned -1 and +1 nucleosomes

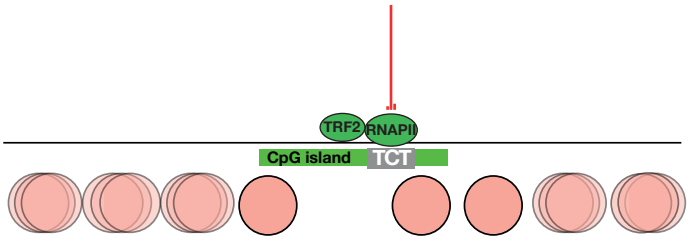


Figure 4

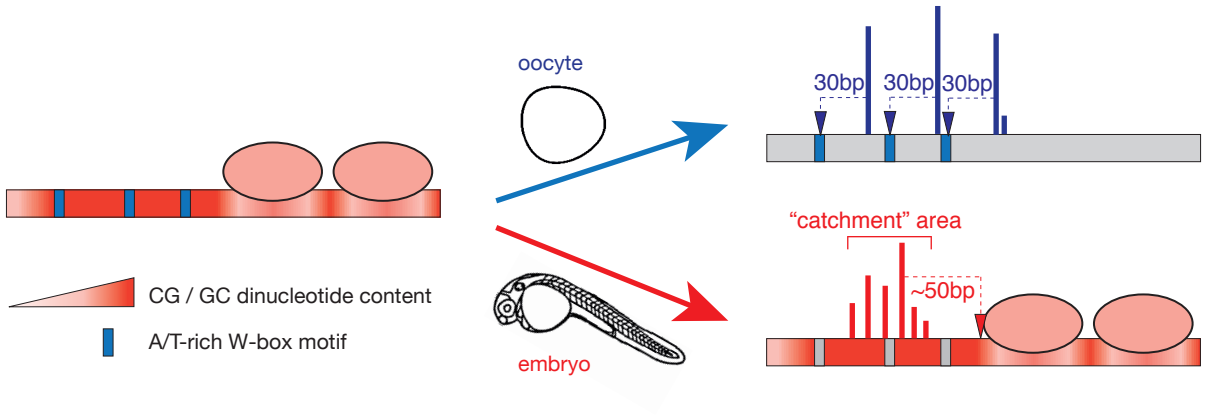
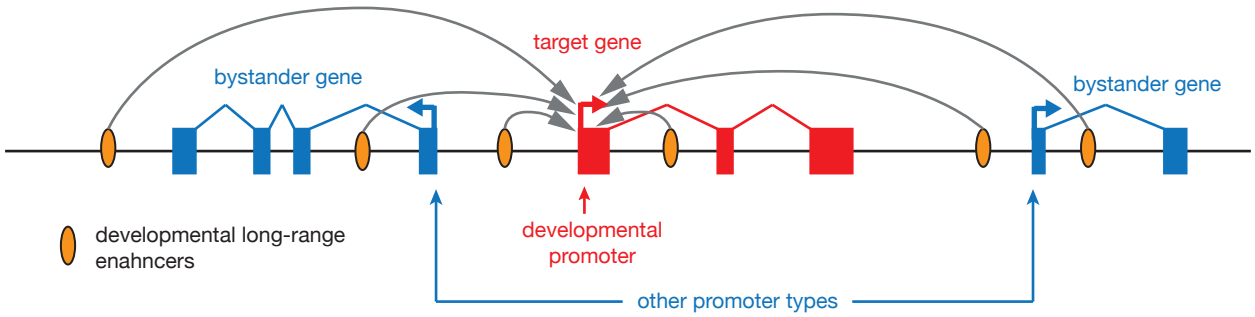

oocyte
30bp
30bp
30bp
CG / GC dinucleotide content
A/T-rich W-box motif
embryo
“catchment” area
~50bp

Figure 5


target gene
bystander gene
bystander gene
developmental long-range
enahncers
developmental
promoter
other promoter types