Mammalian RNA polymerase II core promoters: insights from genome-wide studies

Albin Sandelin[*‡], *Piero Carninci*[‡§], *Boris Lenhard*[‖], *Jasmina Ponjavic*[¶], *Yoshihide Hayashizaki*[‡§] *and David A. Hume*[#]

Abstract | The identification and characterization of mammalian core promoters and transcription start sites is a prerequisite to understanding how RNA polymerase II transcription is controlled. New experimental technologies have enabled genome-wide discovery and characterization of core promoters, revealing that most mammalian genes do not conform to the simple model in which a TATA box directs transcription from a single defined nucleotide position. In fact, most genes have multiple promoters, within which there are multiple start sites, and alternative promoter usage generates diversity and complexity in the mammalian transcriptome and proteome. Promoters can be described by their start site usage distribution, which is coupled to the occurrence of *cis*-regulatory elements, gene function and evolutionary constraints. A comprehensive survey of mammalian promoters is a major step towards describing and understanding transcriptional control networks.

Transcription start site
A nucleotide in the genome that is the first to be transcribed into a particular RNA.

Every aspect of homeostasis, growth, differentiation and development in eukaryotes requires the regulated production of specific mRNAs by RNA polymerase II (RNApolII). The mechanisms that underlie this regulation have been the subject of intense genetic, biochemical and computational studies[1–8]. The transcription start site (TSS) of a gene is the first nucleotide that is copied at the 5′ end of the corresponding mRNA. The region around a TSS is often referred to as the core promoter — which is required for recruitment of the transcription apparatus and can be thought of as the priming stage for transcription initiation. Owing to the strong link between TSSs and core promoters, the terms are often used interchangeably. A confounding issue is that many genes have multiple TSSs that are located in close proximity to each other. For clarity, we define the TSS as a unique nucleotide that will be the first to be transcribed, whereas the core promoter is defined as a genomic region that spans this and the nearby TSSs.

In the past, TSSs and core promoters have been identified on a gene-by-gene basis by one of two methods, nuclease protection or primer extension (BOX 1), neither of which can be applied on a genome-wide scale. Known TSSs are used to find *cis*-regulatory elements that are assumed to lie upstream of the TSS. This is frequently done by fusing the upstream sequence to a reporter gene and then introducing targeted deletions in that sequence to decipher where crucial elements reside. This approach is often referred to as a reporter gene assay, and underlies much of our knowledge of *cis*-regulatory elements.

Cis-regulatory elements of core promoters are commonly conserved across orthologous genes and contribute to the specificity of transcription initiation[1]. However, the complete set of mammalian promoters is too diverse to allow reliable computational annotation of genomic DNA sequence without reference to the experimentally determined locations of full-length cDNA sequences or TSSs from orthologous genes[8,9]. This limitation has motivated the development and application of high-throughput methods to experimentally identify TSSs and their flanking core promoters as a foundation for understanding transcriptional control, and as a tool for genome annotation. Results from these high-throughput studies have revealed a surprisingly large number of novel intergenic transcripts and promoters, prompting us to rethink mammalian promoter architecture.

Here we review recent relevant studies and technological developments to describe the current state of our knowledge of core promoter architecture and function, and highlight future challenges.

[#]*ARC Special Research Centre for Functional and Applied Genomics, Institute for Molecular Bioscience, The University of Queensland, Brisbane QLD, 4072, Australia.*
Correspondence to D.A.H.
e-mail: d.hume@imb.uq.edu.au
doi:10.1038/nrg2026
Published online 8 May 2007

Core promoter
The genomic region that surrounds a TSS or cluster of TSSs. There is no absolute definition for the length of a core promoter; it is generally defined empirically as the segment of DNA that is required to recruit the transcription initiation complex and initiate transcription, given the appropriate external signals (such as enhancers).

Orthologues
Genes that originate from the same ancestral gene and are diverged by a speciation event.

Mediator complex
A multi-subunit complex that can respond to many different activators (such as DNA-bound transcription factors) and links such signals to the core promoter and the transcription machinery.

Tag library
A tag library is similar to a conventional cDNA library, except that, subsequently to isolation and cloning of the cDNA, small fragments are generated by restriction-enzyme cleavage, concatamerized and recloned. This approach enables efficient DNA sequencing of thousands of tags from a single library.

RNApolII transcription initiation

The 'textbook' model of an RNApolII promoter has an AT-rich DNA sequence (the TATA box) approximately 30 bp upstream of an initiator (Inr) sequence that contains the TSS. Assembly of a pre-initiation complex (PIC), which includes the transcription factor TFIIA-H along with RNApolII, at such promoters is initiated by TFIID binding to the TATA box, Inr sequences and/or other sites, and bending DNA through a 90° angle. The next step involves recruitment of general transcription factors[2], after which transcription is initiated 30 bp downstream (see REFS 1–4,10,11 for reviews). Basal and regulated transcription initiation generally also involves interactions of the PIC with three additional components: the TATA-associated factors, the so-called mediator complex(es), and positive and negative cofactors. Coordination of chromatin modification, mainly through the control of post-translational modification of histones, also has an important role in transcription initiation[1–6]. The recruitment of all of these co-activators and co-repressors of transcription initiation is controlled by transcription factor binding to *cis*-acting DNA sequences that can lie within the core promoter or in more remote locations (enhancers and repressors)[8].

Apart from the TATA-box, subsets of promoters contain the Inr element, CpG islands and other sequence patterns[1] (BOX 2), but their prevalence and role in the initiation of transcription are not as well characterized.

It is now clear that TATA-driven PIC assembly is the exception, rather than the rule, in eukaryotic transcription, as only a fraction of mammalian promoters (10–20% (REFS 12,13)) contain a functional TATA box. Similar conclusions have been reached from genome-wide analysis of *Drosophila melanogaster*[14,15] and *Arabidopsis thaliana*[16] promoters. In fact, in mammals, TATA-containing promoters are commonly associated with tissue- or context-specific genes[17].

Genome-wide TSS discovery

The availability of the genome sequences of many eukaryotes has enabled the development of methods designed to analyse features such as gene boundaries, epigenetic effects and active *cis*-regulatory sites on a genome-wide scale[10,11,18–24]. A subset of these are targeted specifically towards locating the 5′ boundaries of transcripts or active TSSs[10,11,12,23–27] (BOX 3). Genome-wide approaches involve several sequencing-based high-throughput methods that require reliable isolation of full-length cDNAs, sequencing of their 5′ ends and mapping of the sequence to a completed genomic DNA sequence. The sequencing stage can use the 5′ ends of cloned full-length cDNA libraries (so-called 5′ ESTs[28,29]), short tags derived from 5′ ends of capped RNAs (CAGE[30] (cap analysis of gene expression; see the CAGE Basic Viewer and the CAGE Analysis Viewer web sites) and 5′-SAGE[30–33] (serial analysis of gene expression)) or tags derived from 5′–3′ ends (so-called paired-end tags (PET)[34]). All of these methods use reverse transcription to generate cDNA, and a full-length copy of the template RNA is not always produced. The key to accuracy is the stringency with which genuine full-length cDNAs are captured before sequencing. An important feature that distinguishes EST and full-length cDNA sequencing from the other tagging methods is the throughput, which is evident in the data sets that are already available. For the mouse, we have around 100,000 sequenced full-length cDNAs, around 1,000,000 5′-end sequences derived from full-length cDNAs, and more than 10,000,000 CAGE and other 5′-end tags. The depth of sequencing that can be achieved with tag methods means that the frequency of tags in an individual library can be used to indicate levels of expression. With new sequencing technology, tag sequencing might well supplant microarrays as a way of analysing gene expression.

In addition to the sequencing-based methods, another, albeit less precise, approach to identifying promoters involves chromatin immunoprecipitation (ChIP)[35,36] of DNA-bound, promoter-associated proteins, where the bound DNA is applied to tiling arrays known as ChIP–chip) (BOX 1). Alternatively, ChIP material can be used as a substrate for tag library construction, and the ChIP products can be analysed by sequencing[37,38].

Putative 5′-end regions can also be identified by hybridizing labelled cDNA or cRNA to high-density genome tiling arrays. Subsequent large-scale bidirectional cDNA synthesis using gene-specific primers is used to extend and validate the 5′ boundaries and identify promoter regions[19,22,39].

Scaled-up 'standard' labour-intensive molecular biology methods such as reporter gene assays and 5′ RACE on individual genes using gene-specific primers[40] are generally not feasible for whole-genome studies. Nevertheless, they have been used successfully for comprehensive screening of promoters in the 1% of the human genome that was chosen for intensive analysis in the first phase of the ENCODE (Encyclopedia of DNA Elements) project[12] (see the UCSC ENCODE web site).

All of the above approaches (the tag-based methods in particular) have identified a large number of TSSs and associated core promoters. Among them, only those that use PET ditags[34] or 5′–3′-end-paired EST reads[23,28,41] provide information about the length of transcripts that are produced from identified 5′ ends.

Author addresses

*The Bioinformatics Centre, Department of Molecular Biology & Biotech Research and Innovation Centre, Ole Maaløes Vej 5, DK-2100 København Ø, Denmark.
‡Genome Exploration Research Group (Genome Network Project Core Group), RIKEN Genomic Sciences Center (GSC), RIKEN Yokohama Institute, 1-7-22, Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan.
§Genome Science Laboratory, Discovery and Research Institute, RIKEN Wako Institute, 2-1 Hirosawa, Wako, Saitama, 351-0198, Japan.
||Computational Biology Unit, Bergen Center for Computational Science, and Sars International Centre for Marine Molecular Biology, University of Bergen, Thormøhlensgate 55, N-5008 Bergen, Norway.
¶MRC Functional Genetics Unit, Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford OX1 3QX, United Kingdom.

Box 1 | **Overview of hybridization-based methods for TSS identification**

There are two types of method for inferring transcription start sites (TSSs): those that are based on sequencing of cDNAs, and those that involve hybridization of RNA or cDNA to DNA probes. Some methods allow for targeting to a specific genomic region or gene, whereas others, by design, must be applied to whole genomes. Examples of hybridization-based methods are given below, listing the distinctive advantages and drawbacks (see BOX 3 for sequence-based methods).

Nuclease protection assay

Nuclease protection methods rely on hybridizing a labelled DNA probe, designed to be complementary to a postulated TSS region, with a source of mRNA, and incubating with a nuclease (often S1 nuclease) that cleaves single-stranded molecules[97]. The length of the protected fragment (inferred from gel electrophoresis) can be used to deduce the length of the mRNA and the corresponding position of the TSS on genomic DNA.

Keywords. Gel-based, low throughput, targeted.

Advantage. This technology is independent of reverse transcriptase reactions.

Disadvantages. Apart from the low-throughput nature of the method, the gel-based interpretation of transcript sizes (and therefore TSS locations) is difficult if there are many TSSs in close proximity. Also, controlling nuclease activity requires appropriate enzyme calibration, and this method requires the use of radioisotopes.

Primer extension

Primer extension analysis uses a labelled primer that is complementary to an internal region of an mRNA that is used for runoff reverse transcription of the mRNA template. The products are typically subjected to partial DNA digestion, following which the resulting labelled fragments are separated on a denaturing polyacrylamide gel and compared with DNA fragments of known size, typically a sequencing ladder. Variants of the primer extension method involve stringent capture of the 5′ end of the mRNA on the basis of the chemical modification (the cap) on the first base of mammalian mRNAs.

Keywords. Gel-based, low throughput, targeted.

Advantages. This method provides the length of the RNA starting from the primer in a similar way to the rapid amplification of cDNA ends (RACE) method but without amplification, and is less laborious than nuclease protection assays.

Disadvantages. It is low throughput and requires detection with radioisotopes. Compared with RACE, this method requires more starting material but shares all its disadvantages.

Tiling arrays

Regardless of the presence of the cap, RNA is randomly primed in order to minimize any possible bias caused by oligo-dT primers. Once double-stranded cDNAs have been synthesized, they are further fragmented with DNase I and extended with terminal nucleotidyl transferase, which adds biotinylated nucleotides. After hybridization to arrays, an antibody against biotin is used to detect signal (see figure).

Keywords. Hybridization-based, whole genome or targeted, often strandless.

Advantages. Tiling arrays provide a snapshot of all the transcribed regions in the genome, not only the 5′ or 3′ ends. This can be obtained in a single experiment at a fraction of the cost of full-length cDNA sequencing. Chips can be customized for chosen regions or can cover the whole non-repetitive part of the genome.

Disadvantages. This techinique must be complemented by RACE or tags/ditags to infer where the edges of exons lie, especially at the 5′ and 3′ ends. It integrates signals from all transcripts in a sample into a single signal, which means that alternative-splicing information and splicing patterns cannot be distinguished. Some platforms cannot distinguish which of the two DNA strands is being transcribed.

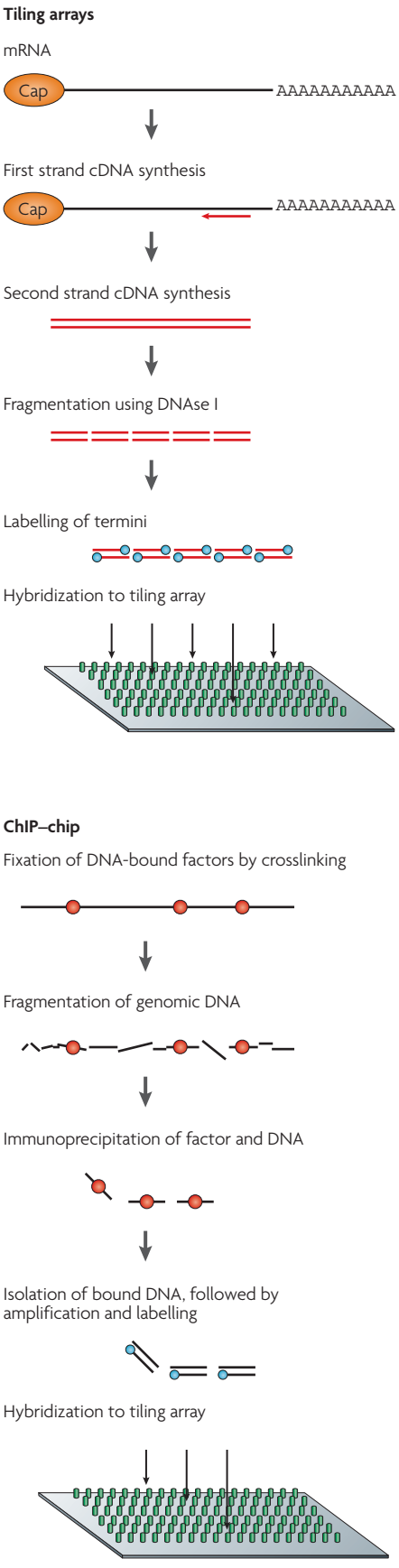
ChIP–chip

In chromatin immunoprecipitation (ChIP), antibodies are used to isolate DNA fragments that are bound to DNA binding proteins or their complexes. After crosslinking the whole proteins to the genomic DNA, the genome is reduced to fragments of a few hundred base pairs in length by sonication. Subsequently, an antibody that recognizes a specific nuclear protein of interest is used to isolate specific complexes. The DNA is then purified, and after ligation of appropriate linkers, it is amplified and labelled. Hybridization of the labelled DNA fragments to whole-genome tiling arrays reveals the genomic location of the DNA (see figure). Alternative protocols have used sequencing instead of DNA arrays.

Keywords: Hybridization-based, provides locations of *in-vivo* DNA-bound proteins

Advantages: ChIP–chip shares many of the features of tiling arrays (see above). However, a unique feature is that biologically active transcription factors or enzymes (such as RNApolII) can be captured 'in action'. Specific antibodies that are targeted to specific molecules can be used (such as phosphorylated RNApolII).

Disadvantages: The hybridization of relatively large DNA fragments after sonication to tiling arrays results in detection of a larger region of DNA than is actually covered by the DNA-bound protein. The results can easily be over-interpreted, as the method shows only the DNA binding site of the protein, but does not necessarily describe the actual function of the interaction.



Box 2 | **Common DNA elements in core promoters**

A set of common DNA sequence elements and patterns are associated with core promoters. These patterns have important characteristics that are linked to the expression of the downstream genes. Different elements can co-occur in the same promoter, although certain combinations are more likely than others, and some patterns complement each other. Below we provide a brief overview of the best studied elements (for a more comprehensive discussion, see REFS 1,2,7). Analysis of the locations of these elements relative to experimentally defined transcription start sites (TSSs) is provided in FIG. 2. For description purposes, the patterns are presented as International Union for Pure and Applied Chemistry (IUPAC) consensus symbols: N, any nucleotide; R, A or G (purine); S, C or G; V, A or C or T; W, A or T; Y, C or T (pyrimidine).

TATA box

The TATA box, located 28–34 bp upstream of the TSS, is perhaps the best known transcription factor binding site. Its consensus sequence, TATAA, binds the TATA-box binding protein (TBP), which is part of the pre-initiation complex (PIC). TATA boxes are associated with strong tissue-specific promoters, and often co-occur with initiator (Inr)-like sequences at the initiation site. Binding of TBP to the TATA box enforces the PIC to select a TSS in a limited genomic space.

Initiator element

The Inr element, defined by the YYANWYY consensus where the A is at position +1 (REF. 1), is independent of the TATA box, although the two can occur together and act synergistically. The TATA and Inr elements are the only known core promoter elements that, alone, can recruit the PIC and initiate transcription.

Downstream promoter element (DPE)

The DPE lies 28–32 bp downstream of the TSS in TATA-less promoters of *Drosophila melanogaster*[15,98]. It has a RGWYV consensus. Generally, it occurs together with Inr elements. The DPE is thought to have a similar function to the TATA box in directing the PIC to a nearby TSS[98].

TFIIB recognition element (BRE)

The BRE element, with an SSRCGCC consensus, lies upstream of the TATA box in some TATA-dependant promoters[99]. It can either increase or decrease transcription rates in eukaryotes, although the details of this meachnism are unknown.

CpG island

CpG islands are genomic stretches in which CG dinucleotides are overrepresentated[100]. On the basis of the original computational definition of CpG islands, 50% of human promoters are associated with CpG islands[101]. Subsequent studies using a statistically derived definition of CpG islands increased this fraction to 72%[102]. CpG-island-associated promoters are most often associated with so-called housekeeping, or ubiquitous, genes[17], although there are many exceptions, including brain-specific genes[103]. Only a fraction of CpG-associated promoters have TATA-like elements.

The most extensive core promoter identification study undertaken so far used CAGE tags to identify 184,379 human and 177,349 mouse core promoters, many of which might contain a cluster of individual TSSs[24]. A previous analysis that involved full-length cDNA sequencing identified 30,964 human and 19,023 mouse promoters[42]. But even the most recent figures are likely to be a substantial underestimate. First, sequencing 50–100,000 tags in each library can reliably detect only those transcripts that are expressed at a level of at least 10 copies in each cell (as there are at least 400,000 mRNAs in an average mammalian cell[43]). Many transcripts are not present at this level, either because they are of low abundance in individual cells or are expressed in only a small subset of cells in the tissues that have been studied.

The second source of underestimation derives from the fact that some CAGE tags cannot be mapped to a single genomic location. A substantial fraction of these map to two or three regions, perhaps indicating that core promoter sub-sequences are shared by distinct promoters. It might be possible to resolve them by merging CAGE data with the extensive collection of 5′ ESTs and PETs. The remaining multi-mapping tags map to repeat regions[24]. There is evidence of promoter activities residing in long terminal repeats of transposons[44,45], but which of these elements are active remains unknown.

The above CAGE-based study focused only on TSSs that were detected more than once (with two or more tags mapping to the same site) and that mapped unequivocally to one genomic location, although many lines of evidence indicated that the large majority of singletons were also true TSSs[24]. The data were validated by various experimental and statistical methods, including RACE, comparison to known full-length cDNA sequences, and conservation of precise TSSs between human and mouse orthologous genes[24]. They are also compatible with independently derived human data sets, which were assembled mainly using oligocapping methods and from full-length cDNA collections that are present in the Database of Transcriptional Start Sites (DBTSS)[42].

Overall, genome-wide approaches, including cDNA sequencing and tiling array studies, reveal a surprisingly large number of novel, intergenic transcripts and promoters[19,22–24]. Many newly identified promoters are not associated with identifiable downstream ORFs, and thus probably direct production of non-protein-coding RNAs. Several novel promoters lie in regions that were thought to be gene deserts on the basis of mapping of full-length cDNAs[19,22–24,39]. This might be correlated with the observation that transcription factor binding sites are often distant from a known gene[18]: there are examples of crucial distal enhancers that are transcribed[46–48], producing non-coding RNA (ncRNA) products that can function in imprinting[46] or as transcriptional co-activators. Given their abundance and apparent diversity, it is thought that many other roles of ncRNAs remain to be discovered.

Box 3 | **Overview of sequencing-based methods for TSS identification**

The methods for inferring transcription start site (TSSs) are either based on sequencing of cDNAs, or rely on hybridization (see BOX 1). Descriptions and distinctive advantages and drawbacks for sequence-based methods are listed below.

RACE

Rapid amplification of cDNA ends (RACE)[40] is used to detect the 5′ ends of individual RNAs. It can be thought of as a 'next-generation' primer extension protocol. RNA is at first oligo-capped; a phosphatase treatment removes the phosphate groups from truncated or uncapped RNA molecules, whereas full-length mRNAs remain protected by the cap structure. Subsequently, the cap is removed by tobacco acid pyrophosphatase, leaving a 5′-end phosphate group that is used by RNA ligase to conjugate an oligonucleotide to the 5′ end. For specific transcripts, primers (or a set of nested primers to increase specificity) are then used for a reverse transcription (RT)-PCR reaction. The product is cloned and sequenced (see figure).

Keywords: Sequencing-based, low throughput, targeted

Advantages: RACE is highly sensitive and complements 5′ tagging. This method is useful for targeting particular loci of interest with higher scalability than hybidization-based approaches, making it ideal for verifying particular TSSs that have been identified by high-throughput methods.

Disadvantages: It is a low-throughput method, as different primers should be designed to validate each TSS in separate experiments. Similar to 5′ tagging methods, information about the span of the full transcript is not retained.

5′ tag sequencing

5′ tag sequencing[25], exemplified by the cap analysis of gene expression (CAGE) technique, allows high-throughput identification of TSSs. Cap-trapping ensures that only the cDNA molecules that are extended to the (biotinylated) cap site are selected; cDNAs that fail to reach the cap site are not selected, because RNase treatment removes the biotinylated cap from the partial cDNA–mRNA hybrid molecule. A linker that contains an *Mme*I restriction site is then ligated to the 5′ end of cDNAs. After the second strand synthesis, *Mme*I cleaves 20–21 nucleotides within the cDNA, producing a 5′-end tag. After ligation with another linker and PCR amplification, tags are concatenated and cloned in a plasmid vector. Concatenation enables sequencing of multiple tags in a single run, decreasing costs (see figure). Sequenced tags are then mapped to the genome using alignment programs such as BLAST[104].

Keywords: Sequencing-based, high throughput, genome-wide only

Advantages: These methods have the highest throughputs. A unique feature is that both TSS location and degree of usage (the number of tags mapping to a specific location is roughly proportional to the transcription level) can be determined. Using material from many different tissues, tissue-specific promoters can be located.

Disadvantages: The short span that is covered by the tags make some alignments unusable (for instance, mappings of transcribed repeat elements). Information about the span of the full transcript is not retained.

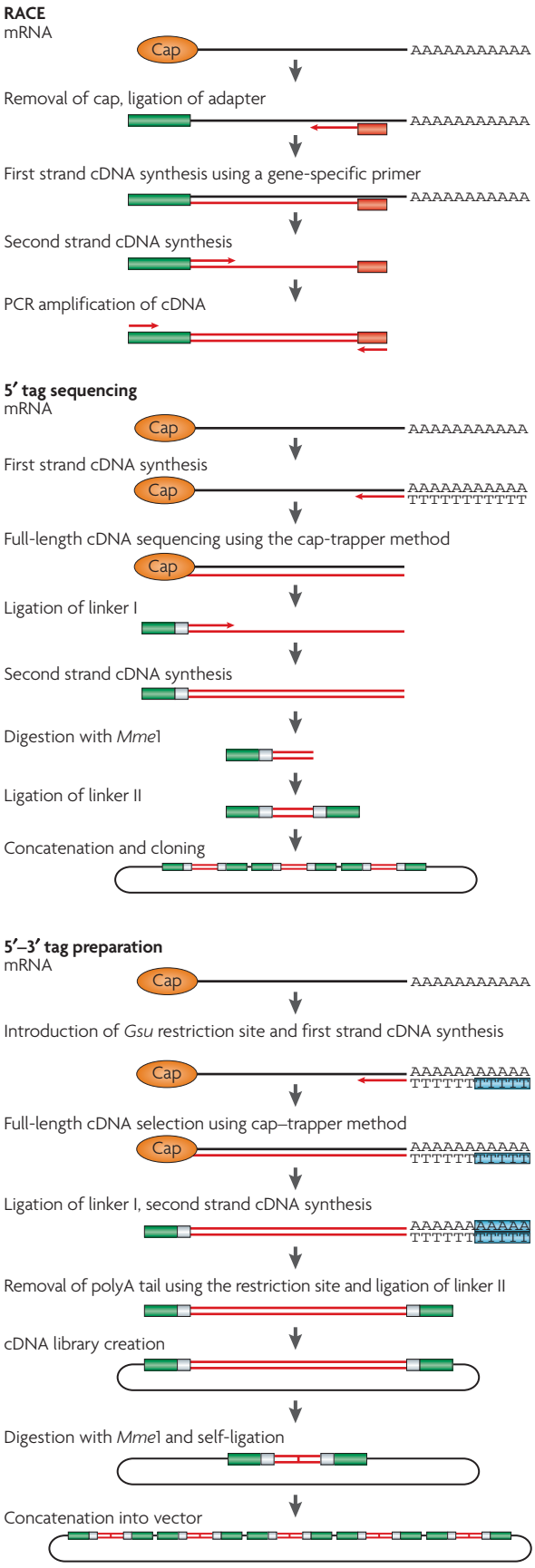
5′–3′ tag sequencing

Exemplified by paired-end ditag technology (PET), a full-length cDNA library is first prepared using the cap-trapping method. An oligo-dT primer is used to prime the first strand cDNA synthesis. The oligo also carries a *Gsu*I restriction site, which enables the removal of the 3′ polyadenylated stretch following restriction with *Gsu*I. The cap-trapped, deadenylated full-length cDNA is ligated to a linker at the 3′ end, which is used for subsequent cloning into a plasmid vector. As there are *Mme*I sites at both cDNA ends, *Mme*I cleavage removes the cDNA insert, except for the 5′–3′-end tags. After re-ligation, the 5′–3′ ditags are excised again and used to produce concatamers (see figure).

Keywords: Sequencing-based, high throughput, genome-wide only

Advantages: Sequencing of both 5′ and 3′ ends makes assigning TSSs to transcripts less problematic. The method gives additional information because the 3′ UTRs can harbour important *cis*-regulatory elements. It can also be used together with ChIP to sequence DNA that is bound by a factor of interest.

Disadvantages: 5′–3′ tag sequencing is a lower-throughput method than 5′-end tagging. As there are more steps in the protocol, the risk of introducing bias increases during the numerous DNA amplification steps, which might cause the information that can be derived from long mRNAs to be lost.



The integration of TSS, transcription factor binding data and phylogenetic information from multiple mammalian genomic sequences has the potential to identify new classes of genome regulatory elements in regions we currently regard as 'intergenic'. Even with our knowledge to date, the conceptual framework of a genome that is simply organized into distinct gene units must be revised[49]. What emerges instead is a transcriptional landscape in which there are no obvious boundaries between the units we currently call genes[23].

Redefining the anatomy of core promoters

The CAGE-based approach indicated that most human and mouse promoters lack the distinct TSS that is commonly assumed to be located at one specific genomic position; instead, the typical core promoter architecture consists of an array of closely located TSSs that spread over around 50–100 bp[23,24]. Many hybrids between these two types of promoter also exist; for instance, in some promoters, TSSs are distributed over a large region, but most transcription initiates at one specific nucleotide position. These observations provide the basis for a new system of promoter classification — it seems more relevant to describe promoters using a TSS distribution that shows the preferred initiation-site usage in a genomic window, instead of a static single position (FIG. 1).

The mapping of smaller datasets of 5′ ends of cDNAs[50] and in-depth studies of individual CpG-enriched promoters (reviewed in REF. 1) have also identified broad TSS regions. Adding further support to this classification, orthologous human and mouse promoters share strikingly similar profiles of TSS usage (FIG. 1). Although a finer subclassification of promoters on the basis of TSS distribution has been proposed[24], for simplicity, we here refer to any of the broad TSS classes as 'broad' and the single TSS groups as 'sharp'. In general, the second classification correlates with the presence of a TATA box, as this feature is associated with promoters that have a single, sharply defined TSS (FIG. 2). The prevalence of distinct TSSs in TATA-box-containing promoters has also been confirmed by analysis of smaller promoter sets[50]. However, not all sharp promoters have a TATA-box (FIGS 1,2), and it will be interesting to examine these cases in detail in future.

Whereas TATA boxes are mostly found in sharp promoters, CpG islands are overrepresented in broad promoters (BOX 2). Sharp promoters are primarily used for tissue-specific expression, whereas broad promoters are generally associated with ubiquitously expressed genes, which is also true for promoters that are associated with TATA boxes and CpG islands, respectively (BOX 2).

Statistical analysis[51] and structural studies[52] indicate that TATA-box position can vary 28–34 bp from the first T in the TATAA consensus to the major initiation-site peak, with a distance of 30–31 bp being strongly preferred. These observations also agree with experimental evidence that shows that, if the TATA site is moved outside these boundaries, new initiation sites that correspond to more favourable TATA–TSS distances arise[53–55]. Ponjavic *et al*[51] showed that the tissue specificity of the promoter is generally highest if the TATA box is located at –30 bp or –31 bp relative to the dominant TSS, and that the initiation-site consensus varies with the TATA–TSS distance. This and another study[16] indicate that longer TATA–TSS distances (32–34 bp) are used more frequently than the shorter ones (28–29 bp). The selection of more distant TSSs seems to be driven by the lack of strong initiation sites (see below) at a more favourable distance.

As noted above, as the number of mammalian promoters that have beeen analysed increases, the proportion of these that contain TATA boxes has decreased. This trend probably reflects the preference among the pioneers of mammalian promoter analysis towards studying highly expressed, tissue-specific genes. This fraction is likely to fall further as more data become available; given the current data coverage, the selection of promoters to study remains biased towards highly expressed genes.

Precise transcription initiation in TATA-box-containing promoters generally requires both the TATA box and an Inr-like element[54]. In a small subset of TATA-less promoters, the Inr sequence alone seems to be able to direct initiation at a single precise location[56,57]. However, most promoters, regardless of class, lack the classical Inr consensus sequence (FIG. 2). Even in its absence, the precise start point of RNApolII-mediated transcription is not random. The [–1,+1] dinucleotide relative to the initiation site shows strong conservation over the whole set of core promoters (a pyrimidine–purine (PyPu) consensus)[24]. The requirement for a purine at +1 is more stringent, but the reported strict requirement for A as a start position[1] is not universal.

Despite the minimal sequence constraint, the importance of the dinucleotide initiator is evident from evolutionary studies. Comparing tag frequencies in orthologous mouse and human promoters reveals that gain or loss of the TSS between species correlates with the creation or removal of the PyPu initiation site[24], respectively. The initiation-site preference varies with initiation-site usage. Highly used TSSs tend to use CG, TG and CA dinucleotides, whereas rarely used TSSs diverge from the preferred PyPu dinucleotide, particularly favouring GG[24]. An example of this can be seen in FIG. 1Bb, in which most of the larger TSS peaks lie over CG, TG or CA sites. The key role of the initiator dinucleotide is supported by a recent study of the ankyrin 1 promoter. In a human patient, deletion of a TG dinucleotide that is used as a TSS in this gene was associated with reduced promoter activity and abolition of TFIID binding to that particular TSS[58].

One problem with using multiple start sites over an extended genomic region is that translation generally starts with the first ATG in an mRNA. Broad promoters must therefore exclude ATG start codons from a certain region, so that all mRNAs that are generated can be efficiently translated. In support of this, a recent study of MHC class I genes identified a set of TATA-less and Inr-less promoters that have multiple TSSs and exhibit such a depletion of ATG trinucleotides in the promoter region[59]. Using orthologous promoters in human, mouse and rat genomes, the authors estimated

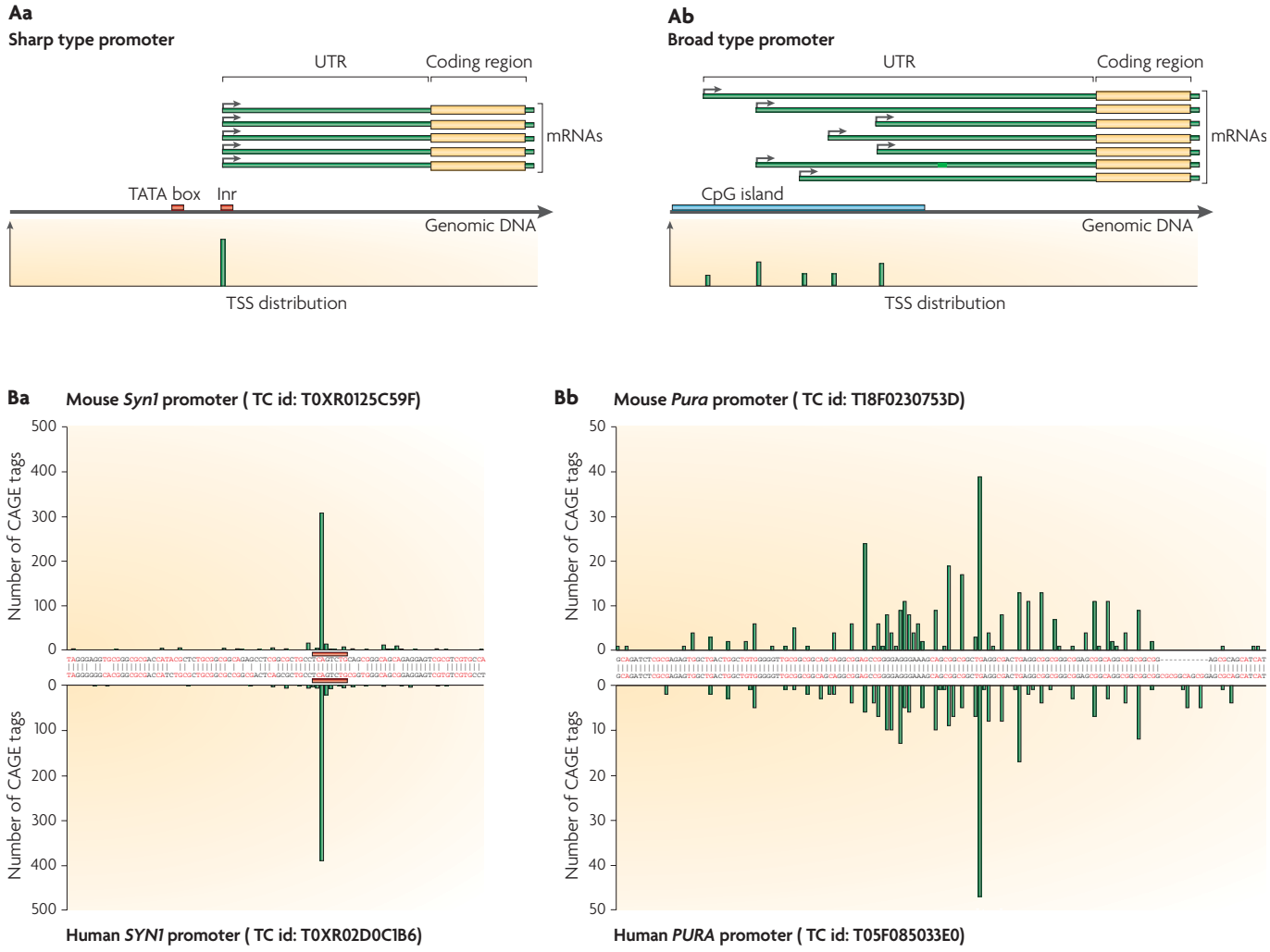


Figure 1 | **TSS classes in mammalian promoters.** Promoters can be classified with respect to the distribution of the transcription start sites (TSSs) they use. For each class, we first show the general features of the class (part **A**) and then a detailed example (part **B**). For these examples, the CAGE (cap analysis of gene expression) tag distribution of the TSSs in the mouse genome (top panel) and of the orthologous TSSs in the human genome (bottom panel) is shown on the *Y* axis. The *X* axis shows the alignment positions between the mouse and human promoter region. Pyrimidine–purine (PyPu) dinucleotides are coloured red — note the correspondence to the major initiation-site peaks. **Aa** | Promoters that fall into the 'sharp' class use only one or a few consecutive nucleotides as TSSs, resulting in a single-peak TSS distribution. These promoters often have TATA and initiator (Inr) boxes. **Ab** | Promoters that fall into the 'broad' class can initiate transcription over a ~100 bp region, resulting in a population of mRNAs that have different lengths but usually the same protein-coding content. Broad promoters are often TATA-less and CpG-island-enriched. **Ba** | The single-peak promoter class is exemplified by the *Syn1* gene, in which almost all TSSs are concentrated in a few consecutive positions, consistent with textbook models of promoters. This particular promoter has a distinct Inr sequence that directs transcription (indicated by a red line), but no obvious TATA box. **Bb** | The *Pura* gene promoter consists of TSSs that are spread over a larger genomic space. Most of the main TSSs in this promoter have a PyPu dinucleotide, but there are no clear Inr consensus sequences. Note the corespondance between human and mouse TSS usage.

that ~82% (6,595 of 8,003) of human genes have 'ATG-desert' characteristics. Collectively, these new data force us to re-evaluate our understanding of transcriptional initiation. The data imply that the TFIID complex, which is essential for all classes of RNApolII transcription, actually binds relatively non-specifically, without an absolute preference for promoters with a TATA box or the strong Inr-like element, and scans along the DNA for a TSS[60].

Tag cluster
This Review defines tag clusters as genomic regions in which two or more tags (of 20 nucleotides in length) overlap each other (both being mapped to the same strand).

The size of core promoters

Studies of TSS usage with genome-scale approaches have provided an overview of the characteristics of broad promoters as a class. The median genomic span that is covered by CAGE tag clusters in broad promoters is 71 bp, and is seldom larger than 150 bp (94.3% of broad promoters that were assessed are smaller than 150 bp when considering the minimum nucleotide range that contains 75% of tags in a cluster).

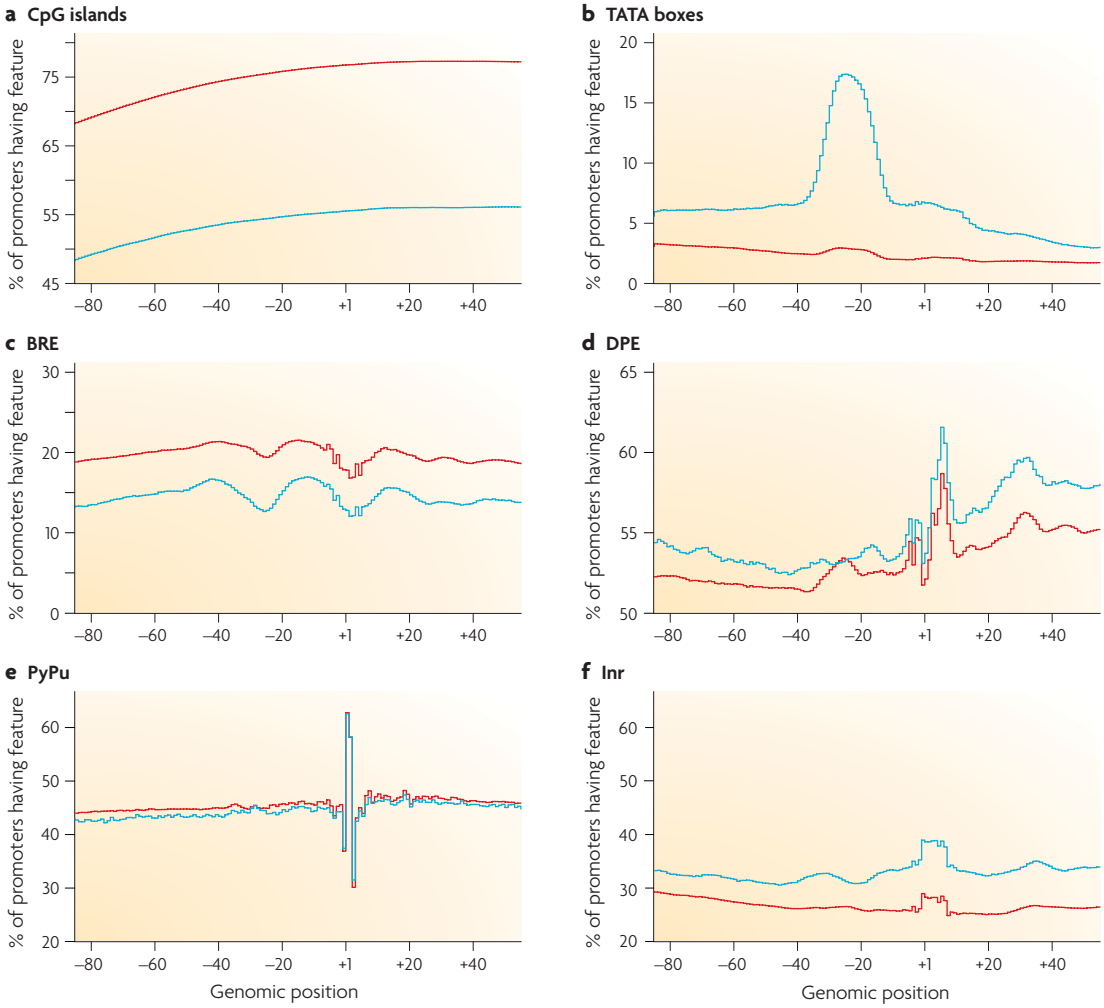


Figure 2 | **DNA motif predictions in core promoters depends on TSS architecture.** Transcription start site (TSS) clusters from a CAGE (cap analysis of gene expression) study[24] that used more than 100 tags from mouse were divided into sharp (indicated in blue) or broad (indicated in red) TSS distribution classes. For each TSS position in the clusters, the –80 to +40 promoter region was scanned using matrix models[8] for core promoter elements and analysed for CpG-island overlap, and the fraction of nucleotides in each position that belonged to the pattern in question was calculated (see Supplementary information S1 (box) for methodology). It is important to note that the absolute frequencies of detected sites are strongly dependent on the cutoffs that are specified in the model, although the aim here is to highlight the contrasts between different types of core promoter. **a** | The broad promoter class has substantially higher CpG coverage than the sharp class, agreeing with previous results[24]. **b** | Consistent with panel **a**, the sharp class has a higher fraction of TATA-matching nucleotides, which are concentrated at the expected location (around –30 to –22). With the cutoff value used here, only around 17% of the sharp class promoters have a canonical TATA site. Even if the fraction was twice as high, this is significantly lower than expected on the basis of the earlier belief that most promoters have a TATA box (see main text); the sharp class of promoters is estimated to cover only about 25% of all promoters. **c** |The BRE (TFIIB recognition) element is reported to occur primarily in TATA-box promoters, just upstream of the TATA box. Surprisingly, this element occurs more often in broad-class promoters (which have significantly fewer TATA sites). This could be due to the higher GC content that is generally observed in broad-class promoters. Although the GC content is increased at the expected location (–37 to –32), a similar increase occurs at the –20 to –1 region. As this increase is modest compared with the background level, it is possible that BRE elements are used less in mammals than in *Drosophila melanogaster*, in which it first was discovered. **d** | The downstream promoter element (DPE) has been reported to lie in the +28 to +32 region in TATA-less promoters in *D. melanogaster*. At this position, the element is frequently observed in both promoter classes, but is most evident in the sharp promoter class, which is consistent with its ability to act as a substitute for the TATA box in directing the precise TSS selection. It is currently unclear whether the DPE is important in mammalian promoters. **e** | The number of pyrimidine–purine (PyPu) dinucleotides is approximately equivalent between the two classes at the actual TSS, but is higher for the broad class in the regions that flank it. This is primarily due to the large number of nearby TSS locations. **f** | The results in part **e** are in sharp contrast to the classical initiator (Inr) element, which is used less often than PyPu dinucleotides in both classes and is most prevalent in the sharp promoter class. The location of the Inr elements is consistent with previous knowledge, indicating that a subset of both classes use the Inr motif instead of the more ambiguous PyPu dinucleotide.

This width constraint is probably related to the length of DNA that is wrapped around a nucleosome and that of the linker DNA, which is ~150 bp[61]. Human promoters have been shown to be nucleosome free at the actual TSS region[62]. If the actual TSSs within such a nucleosome-free region are determined only by the ability of the TFIID complex to bind with some preference to PyPu dinucleotides, as discussed above, the promoter activity of broad-class promoters would be determined solely by the precise position of the nucleosomes, which in turn is regulated by post-translational modification of histones (see REFS 63,64 for reviews). Within these regions, the role of known core promoter elements such as the TATA box would be to restrict the TSS selection process to specific nucleotides.

Interestingly, Segal *et al.*[65] have recently provided evidence for the existence of nucleosome-positioning signals in the vicinity of yeast promoters, suggesting that there is an intrinsic tendency for promoters to exclude nucleosomes. Nevertheless, Kawaji *et al*[66] showed that some broad promoters have overlapping but distinct TSS distributions in different tissues, proving that TSS selection within the promoter can be regulated.

As part of the ENCODE project, a recent study used full-length cDNA mapping and reporter gene assays in 16 cell lines to validate 642 human promoters[12]. The study showed that deletions in the −350 to −40 region upstream of TSSs resulted in decreased reporter gene signals; this region roughly corresponds to the region that is conserved between the promoters of orthologous genes in mice and humans (see REF. 67 and below). The actual crucial interval is probably smaller in any one promoter, as in this study a single arbitrary TSS was chosen as the reference point (so parts of broad TSS regions were probably removed in a subset of cases), and many of the studied intervals contained more than one independently regulated promoter (see below). Focusing on small proximal promoter regions, taking into account the broad TSS regions, should expedite the computational identification of functional motifs that are conserved among classes of genes with common regulatory patterns.

Bidirectional promoters

Previous frameworks for promoter analysis frequently assumed that genes, and therefore their promoters, occupy distinct, non-overlapping genomic regions separated by non-functional DNA. Early studies identified one of the exceptions to this 'rule'[68–70]: genes that lie on opposite strands with their TSSs lying in close proximity to each other form so-called bidirectional promoters. Trinklein *et al*[71]. estimated that 1,352 gene pairs in the human genome have TSSs on the opposite strand that are separated by less than 1 kb; the corresponding number in the mouse was estimated to be 1,638 (REF. 72). Genome-wide analyses that take account of the many newly identified non-coding RNAs and CAGE tags have revealed that promoter overlap of this kind is even more common[23,24,73]. In the large majority of bidirectional promoters, the TSS distribution is of the broad type, although each promoter in a bidirectional pair has independent core promoter elements (the TSS distributions in the two directions generally do not overlap)[24]. In fact, Trinklein *et al*[71] reported that 23% of bidirectional pairs produce sense–antisense pairs that overlap at their 5′ ends, where the TSS region of one gene is upstream of the TSS region of its partner; CAGE data analysis suggests that this is an underestimate[24]. Hence, many bidirectional promoters might be more appropriately referred to as anti-directional or opposing promoter pairs.

Widespread alternative promoter usage

Most mouse and human protein-coding genes are associated with more than one promoter region[12,24,27,42,74]. These alternative promoters are generally used in different contexts or tissues, or to produce distinct protein products. In many cases, the different promoters generate alternative 5′ exons that might or might not contain alternative start codons, and that often splice into a common second exon. The same locus can be associated with both sharp and broad promoters or with multiple broad promoters, each of which has its own ATG desert. For example, the UDP-glucuronosyltransferase locus has at least seven promoters with different tissue expression profiles, each of which produces an alternative first exon and generates a distinct amino (N)-terminal sequence[12,24]. Similarly, the gelsolin gene (*GSN*) can be transcribed to produce a secreted plasma protein that functions as a scavenger of actin filaments or as a cytoplasmic regulator of the cytoskeleton, each from distinct macrophage- or liver-specific promoters, respectively[24]. It is generally accepted that alternative promoter use substantially contributes to the complexity of the mammalian proteome[23,24,73].

Alternative promoters within 3′ UTRs. Most well-supported alternative promoters are found either at the 5′ ends of known cDNAs or in protein-coding exons. However, Carninci *et al.* revealed clusters of TSSs in the last 20% of the terminal exons of protein-coding genes (mostly in 3′ UTRs), on the sense strand of the transcript[23]. At least 1,000 mouse transcription units are associated with at least one such TSS, and the associated core promoters are sufficient to drive transcription[24]. The function of such promoters and their corresponding transcripts are unknown. According to one hypothesis, some of the resulting transcripts will overlap with downstream genes on the other strand, forming potential *cis*-antisense pairs that could contribute to coordinated expression of neighbouring loci[23].

Weak alternative exonic promoters. Brodsky *et al.*[10] analysed the locations of active DNA-bound RNApolII in HeLa cells. Surprisingly, they found that RNApolII sites were concentrated preferentially in exons. The density of RNApolII sites in exons varied between genes, but did not correlate with mRNA levels. The authors attribute these unexpected results to a possible slowdown or pausing of RNApolII elongation within exons[75]. Indeed, an earlier study showed that variations in the speed of RNApolII elongation will affect the usage of splice sites[76]. Accordingly, the amount of exonic RNApolII sites was

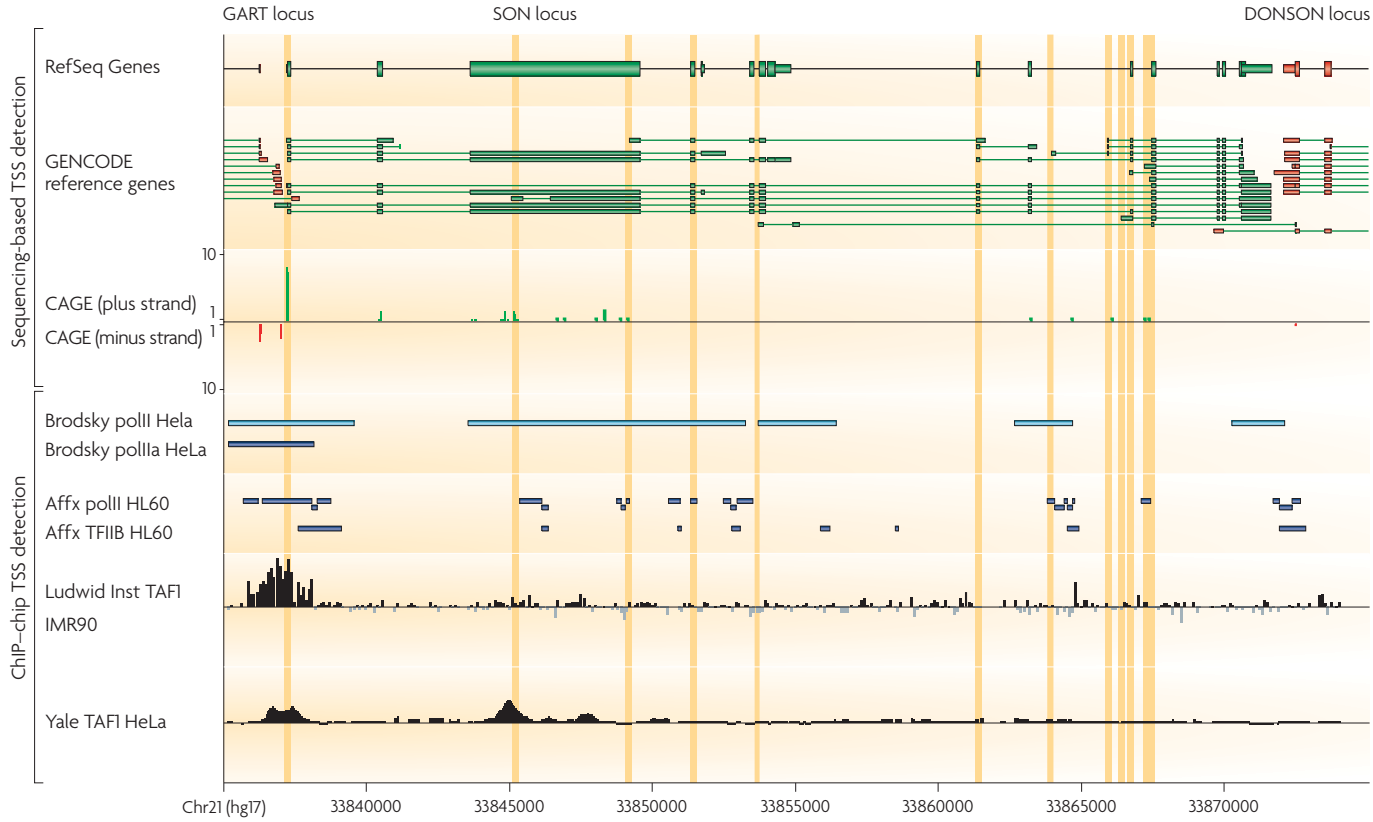


Figure 3 | **Complex TSS distributions within exons.** An example of exonic transcription start sites (TSSs) detected within the ENCODE[105] Enm005 region (human assembly hg17) using multiple techniques, on the basis of the UCSC ENCODE genome browser[96] representation, is shown. The RefSeq[74] track is collapsed owing to space limitations. The strand of the cDNA-based data is indicated in red (reverse) and green (forward). The GENCODE[107] track represents high-quality manual gene annotations from cDNA and EST sources, validated by RACE (rapid amplification of cDNA ends). TSSs within the SON locus, as indicated by GENCODE annotations, are highlighted with yellow columns. The results of CAGE (cap analysis of gene expression) tag sequencing and four distinct chromatin immunoprecipitation (ChIP)–chip experiments are shown as separate tracks below. All ChIP experiments except the Brodsky polII track were made using antibodies targeted to the pre-initiation complex. Many TSSs exist within RefSeq exons, as indicated by CAGE, GENCODE annotation and ChIP–chip. The positions of the majority of GENCODE TSSs that are located within RefSeq exons are supported by one or more of the other technologies; moreover, CAGE data indicate there are additional TSSs that are not detected by GENCODE annotation.

greater in alternatively spliced exons compared with the invariantly spliced ones[10].

A more provocative, although not mutually exclusive, explanation is that many of these RNApolII signals are due to genuine but infrequent initiation events, indicating that there are TSSs scattered within exons. In fact, many low-intensity exonic TSSs were found using CAGE tags[24]. The number of exonic TSSs varies between genes, (for example, exonic TSSs tend to be more prevalent in tissue-specific genes[24]) and this level is conserved between human and mouse orthologous genes. These observations suggest that truncated internally initiated mRNAs constitute a significant class of non-coding mRNAs; this is also consistent with evidence that the initiation complex can bind to sequences within exons but not introns[11] (FIG. 3). The possible function of weak initiation sites within internal exons requires further study, but it could contribute to the recently described phenomenon of exon-tethering — a physical connection between emergent splice sites in pre-mRNA and the RNApolII transcription complex[77].

Evolution of core promoters

Mapping of TSSs to a base-pair resolution on a global scale, together with reliable genome-wide alignments[78], have opened new avenues in promoter evolution studies. Promoter evolution among primate species occurs rapidly through accumulated substitutions and deletions[67,79]. An in-depth comparative evolutionary analysis between primate, mouse, rat and dog core promoters[67] revealed that the substitution rate at each site is lowest in the –50 bp to –1 bp region relative to the dominant TSS, and increases linearly until around –200 bp.

On the basis of the mouse–human comparison, TATA-box-containing promoters evolve more slowly than CpG-island-containing promoters[67]. This finding agrees with an earlier study that showed that context-specific genes have more conserved promoter regions than

other genes[80]. The implication is that the constrained and precise architecture of the TATA-box promoter is needed to ensure reliable transcription initiation in time and space; any change in the functional promoter sequence is likely to have a strong phenotypic consequence. In keeping with this view, a recent study of four closely related yeast species showed that TATA-box-containing promoters have a higher expression profile divergence between species than other promoters[81]; this was also observed in mammals, insects and plants[81]. Slower evolution at the sequence level of TATA-box-containing promoters but faster evolution in expression profiles between species is counterintuitive, but might reflect the fact that such promoters have only one TSS, making the initiation rate much more sensitive to any changes. By contrast, the broad core promoters with their multiple redundant TSSs respond to most mutations by smaller changes in expression, enabling fine-tuning of their activity and facilitating both adaptive evolution and compensation for mutations in the unfavourable direction[24].

The future of promoter analysis

Genome-wide analyses have identified the TATA-less promoter with multiple TSSs as the major class of mammalian promoters, clarified the role of the initiation site, identified sequence constraints in promoter regions, and facilitated more focused studies on the evolution and function of specific proximal promoter elements. They have also identified new classes of promoters that lie within exons and 3′ UTRs, and revealed the widespread use of alternative promoters in protein-coding loci and the functional organization of bidirectional promoter regions. These findings should drive future studies into the precise biochemical mechanisms of transcription initiation in the absence of the TATA box.

The precise location of TSSs enables a focused analysis of *cis*-acting elements that are bound by transcription factors. For example, Carninci *et al.*[24] were able to segregate promoters on the basis of their relative use in different tissues and cell types, and to show that, as would be expected, tissue-specific promoters are enriched for particular motifs that serve as the binding sites for known tissue-specific transcription factors. This analysis provides a starting point for unravelling the molecular details of cooperative interactions among transcription factors[82].

An important future challenge will be to reliably integrate TSS location data with related functional data, such as histone methylation and acetylation states[83,84], the position of nucleosomes[62,85] and the occupancy of transcription factor binding sites[38,86], each of which can now be contemplated on a genome-wide basis[36]. The linking of transcription factor binding sites that are not proximal to any gene with TSS selection for a particular promoter will present another important challenge.

It remains difficult to associate the 5′ ends of genes with their corresponding transcripts, especially if the TSS is novel. Some intragenic TSSs produce transcripts that bridge two or more downstream genes[87] but it might not be biologically relevant to associate these TSSs with the protein-coding gene in which they occur for other non-standard promoter types (3′ UTR promoters in particular).

If there are no cDNAs in a sample that map to the region in which a TSS lies, tiling array data might help to detect nearby transcribed regions (presumed to be exons). However, there is no reliable way of assigning an exon structure or TSSs to transcribed fragments that have been identified by tiling arrays without using full-length cDNAs or extensive cloning of 5′–3′ RACE products.

Chromosome conformation capture is a promising method for linking TSSs and regulatory elements to their corresponding transcript. This technique detects the physical interactions between chromosomal regions that are involved in common regulatory mechanisms[88–90]. This type of data integration is necessary if we are to understand the interaction between regulatory elements and regulatory proteins, and how they function together to direct transcription. Large-scale TSS data is crucial but not sufficient in itself for making this leap.

On a functional level, the next generation of experiments should consider the extent and function of novel transcripts in normal cells, instead of the actively proliferating cells that most ChIP–chip and tiling array data currently come from. Because proliferation requires specific transcriptional programmes, many observed transcripts might encode molecules that are specific to these processes. Despite the success of the genome-wide studies so far, the coverage of TSSs is incomplete; similar methods should be applied to other species, as only two mammalian species (mice and humans) have been investigated in any detail. High-throughput studies should also be extended to different tissues, as properties of promoters vary substantially between different contexts, tissues and cell types.

The multitude of newly found promoters is an obvious boon for computational biologists. The tag-based methods in particular provide both quantitative and nucleotide-positional information on tissue specificity, thus removing the most serious bottleneck for defining tissue-specific *cis*-regulatory elements[91].

For experimental biologists, the genome-wide data sets provide a valuable map of the promoter regions of genes of interest to guide further studies. The depth of data for mice and humans allows functional alignment of promoters[92] in phylogenetic studies as more complete mammalian genome sequences become available. As both tag and tiling array approaches and their related technologies reach maturity, we can expect the generation of genome-wide data sets to increase rapidly; for example, massive parallel sequencing technologies[21,93,94] are being integrated into pipelines for tag sequencing, and new applications of tiling arrays, such as hypersensitive site detection, are appearing[20,95]. The increasing depth of data to describe promoter architecture and diversity leads to numerous hypotheses about the function of motifs and motif combinations that must be tested experimentally. The pilot phase of the ENCODE project[96] and its extension to a larger part of the human genome will be a driving force for assessing new technologies and deepening our understanding of the transcriptome and its regulation.

Conclusions

Large-scale studies re-emphasize the take-home message that has emerged from both old and recent reviews of promoter function, and that has been largely ignored in computational studies: mammalian genomes harbour many types of core promoter. It is now evident that the rules for start site selection are fundamentally different for different promoters, and large-scale studies have given us the tools to partition promoters into functional classes for further study. The use of biologically relevant promoter classification will be vital for the identification of important regulatory signals and mechanisms. Such an approach should eliminate much of the noise that has resulted from indiscriminate mixing of different functional classes in genome-wide analyses, thereby conferring a clear advantage to computational studies.

1. Smale, S. T. & Kadonaga, J. T. The RNA polymerase II core promoter. *Annu. Rev. Biochem.* **72**, 449–479 (2003).
An excellent in-depth review of well-studied core promoter elements.
2. Gross, P. & Oelgeschlager, T. Core promoter-selective RNA polymerase II transcription. *Biochem. Soc. Symp.* **2006**, 225–236 (2006).
3. Hampsey, M. Molecular genetics of the RNA polymerase II general transcriptional machinery. *Microbiol. Mol. Biol. Rev.* **62**, 465–503 (1998).
4. Thomas, M. C. & Chiang, C. M. The general transcription machinery and general cofactors. *Crit. Rev. Biochem. Mol. Biol.* **41**, 105–178 (2006).
5. Lewis, B. A. & Reinberg, D. The mediator coactivator complex: functional and physical roles in transcriptional regulation. *J. Cell Sci.* **116**, 3667–3675 (2003).
6. Black, J. C., Choi, J. E., Lombardo, S. R. & Carey, M. A mechanism for coordinating chromatin modification and preinitiation complex assembly. *Mol. Cell* **23**, 809–818 (2006).
7. Kadonaga, J. T. Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. *Cell* **116**, 247–257 (2004).
8. Wasserman, W. W. & Sandelin, A. Applied Bioinformatics for the identification of regulatory elements. *Nature Rev. Genet.* **5**, 276–287 (2004).
Reviews the computational methods that underlie the prediction of promoter positions and transcription factor binding sites, targeted towards bench biologists.
9. Bajic, V. B., Tan, S. L., Suzuki, Y. & Sugano, S. Promoter prediction analysis on the whole human genome. *Nature Biotechnol.* **22**, 1467–1473 (2004).
10. Brodsky, A. S. *et al.* Genomic mapping of RNA polymerase II reveals sites of co-transcriptional regulation in human cells. *Genome Biol.* **6**, R64 (2005).
This study revealed the surprisingly high concentration of RNApolII that is bound to exons but not introns.
11. Kim, T. H. *et al.* A high-resolution map of active promoters in the human genome. *Nature* **436**, 876–880 (2005).
The first genome-wide ChIP–chip determination using antibodies that targeted the PIC.
12. Cooper, S. J., Trinklein, N. D., Anton, E. D., Nguyen, L. & Myers, R. M. Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Res.* **16**, 1–10 (2006).
13. Gershenzon, N. I. & Ioshikhes, I. P. Synergy of human Pol II core promoter elements revealed by statistical sequence analysis. Bioinformatics **21**, 1295–1300 (2005).
14. 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).
15. Ohler, U., Liao, G. C., Niemann, H. & Rubin, G. M. Computational analysis of core promoters in the *Drosophila* genome. *Genome Biol.* **3**, RESEARCH0087 (2002).
16. Molina, C. & Grotewold, E. Genome wide analysis of *Arabidopsis* core promoters. *BMC Genomics* **6**, 25 (2005).
17. Schug, J. *et al.* Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol.* **6**, R33 (2005).
18. Cawley, S. *et al.* Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**, 499–509 (2004).
19. Cheng, J. *et al.* Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308**, 1149–1154 (2005).
20. Crawford, G. E. *et al.* DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nature Methods* **3**, 503–509 (2006).
21. Fan, J. B., Chee, M. S. & Gunderson, K. L. Highly parallel genomic assays. *Nature Rev. Genet.* **7**, 632–644 (2006).
22. Kapranov, P. *et al.* Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**, 916–919 (2002).
One of several key publications from Affymetrix on the utility of tiling arrays and the widespread occurrence of non-coding RNA.
23. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–1563 (2005).
24. Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genet.* (2006).
The largest experimental promoter identification study to date in any species, with subsequent computational analysis.
25. Harbers, M. & Carninci, P. Tag-based approaches for transcriptome research and genome annotation. *Nature Methods* **2**, 495–502 (2005).
26. Barrera, L. O. & Ren, B. The transcriptional regulatory code of eukaryotic cells — insights from genome-wide analysis of chromatin organization and transcription factor binding. *Curr. Opin. Cell Biol.* **18**, 291–298 (2006).
27. Kimura, K. *et al.* Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res.* **16**, 55–65 (2006).
28. Carninci, P. *et al.* Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia. *Genome Res.* **13**, 1273–1289 (2003).
29. Suzuki, Y. *et al.* Large-scale collection and characterization of promoters of human and mouse genes. *In silico Biol.* **4**, 0036 (2004).
30. Shiraki, T. *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl Acad. Sci. USA* **100**, 15776–15781 (2003).
31. Kodzius, R. *et al.* CAGE: cap analysis of gene expression. *Nature Methods* **3**, 211–222 (2006).
32. Hashimoto, S. *et al.* 5′-end SAGE for the analysis of transcriptional start sites. *Nature Biotechnol.* **22**, 1146–1149 (2004).
33. Wei, C. L. *et al.* 5′ long serial analysis of gene expression (LongSAGE) and 3′ LongSAGE for transcriptome characterization and genome annotation. *Proc. Natl Acad. Sci. USA* **101**, 11701–11706 (2004).
34. Ng, P. *et al.* Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nature Methods* **2**, 105–111 (2005).
35. Shannon, M. F. & Rao, S. Transcription. Of chips and ChIPs. *Science* **296**, 666–669 (2002).
36. Ren, B. & Dynlacht, B. D. Use of chromatin immunoprecipitation assays in genome-wide location analysis of mammalian transcription factors. *Methods Enzymol.* **376**, 304–315 (2004).
37. Loh, Y. H. *et al.* The OCT4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nature Genet.* **38**, 431–440 (2006).
38. Wei, C. L. *et al.* A global map of p53 transcription-factor binding sites in the human genome. *Cell* **124**, 207–219 (2006).
39. Kampa, D. *et al.* Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.* **14**, 331–342 (2004).
40. Schaefer, B. C. Revolutions in rapid amplification of cDNA ends: new strategies for polymerase chain reaction cloning of full-length cDNA ends. *Anal. Biochem.* **227**, 255–273 (1995).
41. Okazaki, Y. *et al.* Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563–573 (2002).
42. Yamashita, R. *et al.* DBTSS: DataBase of Human Transcription Start Sites, progress report 2006. *Nucleic Acids Res.* **34**, D86–D89 (2006).
43. Jackson, D. A., Pombo, A. & Iborra, F. The balance sheet for transcription: an analysis of nuclear RNA metabolism in mammalian cells. *Faseb J.* **14**, 242–254 (2000).
44. Kovalskaya, E., Buzdin, A., Gogvadze, E., Vinogradova, T. & Sverdlov, E. Functional human endogenous retroviral LTR transcription start sites are located between the R and U5 regions. *Virology* **346**, 373–378 (2006).
45. Buzdin, A., Kovalskaya-Alexandrova, E., Gogvadze, E. & Sverdlov, E. GREM, a technique for genome-wide isolation and quantitative analysis of promoter active repeats. *Nucleic Acids Res.* **34**, e67 (2006).
46. Ling, J., Baibakov, B., Pi, W., Emerson, B. M. & Tuan, D. The HS2 enhancer of the β-globin locus control region initiates synthesis of non-coding, polyadenylated RNAs independent of a *cis*-linked globin promoter. *J. Mol. Biol.* **350**, 883–896 (2005).
47. Drewell, R. A. *et al.* Novel conserved elements upstream of the *H19* gene are transcribed and act as mesodermal enhancers. *Development* **129**, 1205–1213 (2002).
48. Ravasi, T. & Hume, D. A. in *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics* (ed. Subramamiam, S.) (John Wiley & Sons, Chichester, 2005).
49. Gingeras, T. R. The multitasking genome. *Nature Genet.* **38**, 608–609 (2006).
50. Suzuki, Y. *et al.* Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. *EMBO Rep.* **2**, 388–393 (2001).
51. Ponjavic, J. *et al.* Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters. *Genome Biol.* **7**, R78 (2006).
52. Hahn, S. Structure and mechanism of the RNA polymerase II transcription machinery. *Nature Struct. Mol. Biol.* **11**, 394–403 (2004).
53. Zhu, Q., Dabi, T. & Lamb, C. TATA box and initiator functions in the accurate transcription of a plant minimal promoter *in vitro*. *Plant Cell* **7**, 1681–1689 (1995).
54. O'Shea-Greenfield, A. & Smale, S. T. Roles of TATA and initiator elements in determining the start site location and direction of RNA polymerase II transcription. *J. Biol. Chem.* **267**, 1391–1402 (1992).
55. Grace, M. L., Chandrasekharan, M. B., Hall, T. C. & Crowe, A. J. Sequence and spacing of TATA box elements are critical for accurate initiation from the β-phaseolin promoter. *J. Biol. Chem.* **279**, 8102–8110 (2004).
56. Smale, S. T. *et al.* The initiator element: a paradigm for core promoter heterogeneity within metazoan protein-coding genes. *Cold Spring Harb. Symp. Quant. Biol.* **63**, 21–31 (1998).
57. Weis, L. & Reinberg, D. Accurate positioning of RNA polymerase II on a natural TATA-less promoter is independent of TATA-binding-protein-associated factors and initiator-binding proteins. *Mol. Cell. Biol.* **17**, 2973–2984 (1997).
58. Gallagher, P. G. *et al.* A dinucleotide deletion in the ankyrin promoter alters gene expression, transcription initiation and TFIID complex formation in hereditary spherocytosis. *Hum. Mol. Genet.* **14**, 2501–2509 (2005).

59. Lee, M. P. *et al.* ATG deserts define a novel core promoter subclass. *Genome Res.* **15**, 1189–1197 (2005).
An in-depth experimental study of promoters with multiple start sites, followed by a computational screening of ATG deserts in the human genome.
60. Carcamo, J., Buckbinder, L. & Reinberg, D. The initiator directs the assembly of a transcription factor IID-dependent transcription complex. *Proc. Natl Acad. Sci. USA* **88**, 8052–8056 (1991).
61. Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251–260 (1997).
62. Nishida, H. *et al.* Histone H3 acetylated at lysine 9 in promoter is associated with low nucleosome density in the vicinity of transcription start site in human cell. *Chromosome Res.* **14**, 203–211 (2006).
63. Mellor, J. Dynamic nucleosomes and gene transcription. *Trends Genet.* **22**, 320–329 (2006).
64. Bantignies, F. & Cavalli, G. Cellular memory and dynamic regulation of polycomb group proteins. *Curr. Opin. Cell Biol.* **18**, 275–283 (2006).
65. Segal, E. *et al.* A genomic code for nucleosome positioning. *Nature* **442**, 772–778 (2006).
66. Kawaji, H. *et al.* Dynamic usage of transcription start sites within core promoters. *Genome Biol.* **7**, R118 (2006).
67. Taylor, M. S. *et al.* Heterotachy in mammalian promoter evolution. *PLoS Genet.* **2**, e30 (2006).
The most comprehensive study of promoter evolution in mammalian species to date.
68. Albig, W., Kioschis, P., Poustka, A., Meergans, K. & Doenecke, D. Human histone gene organization: nonregular arrangement within a large cluster. *Genomics* **40**, 314–322 (1997).
69. Guarguaglini, G. *et al.* Expression of the murine *RanBP1* and *Htf9-c* genes is regulated from a shared bidirectional promoter during cell cycle progression. *Biochem. J.* **325**, 277–286 (1997).
70. Sugimoto, M., Oohashi, T. & Ninomiya, Y. The genes *COL4A5* and *COL4A6*, coding for basement membrane collagen chains {alpha}5(IV) and {alpha}6(IV), are located head-to-head in close proximity on human chromosome Xq22 and COL4A6 is transcribed from two alternative promoters. *Proc. Natl Acad. Sci. USA* **91**, 11679–11683 (1994).
71. Trinklein, N. D. *et al.* An abundance of bidirectional promoters in the human genome. *Genome Res.* **14**, 62–66 (2004).
72. Engstrom, P. G. *et al.* Complex loci in human and mouse genomes. *PLoS Genet.* **2**, e47 (2006).
73. Katayama, S. *et al.* Antisense transcription in the mammalian transcriptome. Science **309**, 1564–1566 (2005).
74. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33**, D501–D504 (2005).
75. Bai, L., Santangelo, T. J. & Wang, M. D. Single-molecule analysis of RNA polymerase transcription. *Annu. Rev. Biophys. Biomol. Struct.* **35**, 343–360 (2006).
76. Kornblihtt, A. R., de la Mata, M., Fededa, J. P., Munoz, M. J. & Nogues, G. Multiple links between transcription and splicing. *RNA* **10**, 1489–1498 (2004).
An excellent review that connects the splicing process to transcription.
77. Dye, M. J., Gromak, N. & Proudfoot, N. J. Exon tethering in transcription by RNA polymerase II. *Mol. Cell* **21**, 849–859 (2006).
78. Schwartz, S. *et al.* Human–mouse alignments with BLASTZ. *Genome Res.* **13**, 103–107 (2003).
79. Keightley, P. D., Lercher, M. J. & Eyre-Walker, A. Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS. Biol.* **3**, e42 (2005).
80. Lee, S., Kohane, I. & Kasif, S. Genes involved in complex adaptive processes tend to have highly conserved upstream regions in mammalian genomes. *BMC Genomics* **6**, 168 (2005).
81. Tirosh, I., Weinberger, A., Carmi, M. & Barkai, N. A genetic signature of interspecies variations in gene expression. *Nature Genet.* **38**, 830–834 (2006).
82. Nilsson, R. *et al.* Transcriptional network dynamics in macrophage activation. Genomics **88**, 133–142 (2006).
83. Yan, C. & Boyd, D. D. Histone H3 acetylation and H3 K4 methylation define distinct chromatin regions permissive for transgene expression. *Mol. Cell. Biol.* **26**, 6357–6371 (2006).
84. Pokholok, D. K. *et al.* Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* **122**, 517–527 (2005).
85. Wiren, M. *et al.* Genomewide analysis of nucleosome density histone acetylation and HDAC function in fission yeast. *EMBO J.* **24**, 2906–2918 (2005).
86. Guccione, E. *et al.* Myc-binding-site recognition in the human genome is determined by chromatin context. *Nature Cell Biol.* **8**, 764–770 (2006).
87. Furuno, M. *et al.* Clusters of internally primed transcripts reveal novel long noncoding RNAs. *PLoS Genet.* **2**, e37 (2006).
88. Wurtele, H. & Chartrand, P. Genome-wide scanning of *HoxB1*-associated loci in mouse ES cells using an open-ended chromosome conformation capture methodology. *Chromosome Res.* **14**, 477–495 (2006).
89. Murrell, A., Heeson, S. & Reik, W. Interaction between differentially methylated regions partitions the imprinted genes *Igf2* and *H19* into parent-specific chromatin loops. *Nature Genet.* **36**, 889–893 (2004).
90. Chakalova, L., Debrand, E., Mitchell, J. A., Osborne, C. S. & Fraser, P. Replication and transcription: shaping the landscape of the genome. *Nature Rev. Genet* **6**, 669–677 (2005).
91. Krivan, W. & Wasserman, W. W. A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res.* **11**, 1559–1566 (2001).
92. Lenhard, B. *et al.* Identification of conserved regulatory elements by comparative genome analysis. *J. Biol.* **2**, 13 (2003).
93. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
94. Ng, P. *et al.* Multiplex sequencing of paired-end ditags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes. *Nucleic Acids Res.* **34**, e84 (2006).
95. Sabo, P. J. *et al.* Genome-scale mapping of DNase I sensitivity *in vivo* using tiling DNA microarrays. *Nature Methods* **3**, 511–518 (2006).
96. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636–640 (2004).
97. Sambrook, J. & Russel, D. W. *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 2001).
98. Kadonaga, J. T. The DPE, a core promoter element for transcription by RNA polymerase II. *Exp. Mol. Med.* **34**, 259–264 (2002).
99. Lagrange, T., Kapanidis, A. N., Tang, H., Reinberg, D. & Ebright, R. H. New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. *Genes Dev.* **12**, 34–44 (1998).
100. Gardiner-Garden, M. & Frommer, M. CpG islands in vertebrate genomes. *J. Mol. Biol.* **196**, 261–282 (1987).
101. Antequera, F. & Bird, A. Number of CpG islands and genes in human and mouse. *Proc. Natl Acad. Sci. USA* **90**, 11995–11999 (1993).
102. Saxonov, S., Berg, P. & Brutlag, D. L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl Acad. Sci. USA* **103**, 1412–1417 (2006).
103. Gustincich, S. *et al.* The complexity of the mammalian transcriptome. *J. Physiol.* **575**, 321–332 (2006).
104. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
105. Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**, D590–D598 (2006).
106. Harrow, J. *et al.* GENCODE: producing a reference annotation for ENCODE. *Genome Biol.* **7**, S41–S49 (2006).

Competing interests statement

The authors declare no competing financial interests.

DATABASES

The following terms in this article are linked online to:
Entrez Gene: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
ankyrin 1 | gelsolin | *Pura* | *Syn1*

FURTHER INFORMATION

CAGE Basic Viewer: http://gerg01.gsc.riken.jp/cage
CAGE Analysis Viewer: http://gerg01.gsc.riken.jp/cage_analysis
FANTOM3: http://fantom3.gsc.riken.jp
Database of Transcriptional Start Sites (DBTSS): http://dbtss.hgc.jp
Eukaryotic promoter database (EPD): http://www.epd.isb-sib.ch
UCSC ENCODE browser: http://genome.ucsc.edu/ENCODE

SUPPLEMENTARY INFORMATION

See online article: S1 (box)
Access to this links box is available online.