Published in final edited form as:
Gene. 2007 March 1; 389(1): 52–65.

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

Chuhu Yang[a,f], **Eugene Bolotin**[a], **Tao Jiang**[b,e], **Frances M. Sladek**[c,e,‡], and **Ernest Martinez**[d,e,‡,*]

[a]*Genetics Genomics and Bioinformatics Graduate Program, University of California, Riverside, CA 92521, USA*

[b]*Department of Computer Science and Engineering, University of California, Riverside, CA 92521, USA*

[c]*Department of Cell Biology and Neuroscience, University of California, Riverside, CA 92521, USA*

[d]*Department of Biochemistry, University of California, Riverside, CA 92521, USA*

[e]*Institute of Integrative Genome Biology, University of California, Riverside, CA 92521, USA*

Abstract

The core promoter of eukaryotic genes is the minimal DNA region that recruits the basal transcription machinery to direct efficient and accurate transcription initiation. The fraction of human and yeast genes that contain specific core promoter elements such as the TATA box and the initiator (INR) remains unclear and core promoter motifs specific for TATA-less genes remain to be identified. Here, we present genome-scale computational analyses indicating that ~76% of human core promoters lack TATA-like elements, have a high GC content, and are enriched in Sp1 binding sites. We further identify two motifs - M3 (SCGGAAGY) and M22 (TGCGCANK) - that occur preferentially in human TATA-less core promoters. About 24% of human genes have a TATA-like element and their promoters are generally AT-rich; however, only ~10% of these TATA-containing promoters have the canonical TATA box (TATAWAWR). In contrast, ~46% of human core promoters contain the consensus INR (YYANWYY) and ~30% are INR-containing TATA-less genes. Significantly, ~46% of human promoters lack both TATA-like and consensus INR elements. Surprisingly, mammalian-type INR sequences are present - and tend to cluster - in the transcription start site (TSS) region of ~40% of yeast core promoters and the frequency of specific core promoter types appears to be conserved in yeast and human genomes. Gene Ontology analyses reveal that TATA-less genes in humans, as in yeast, are frequently involved in basic “housekeeping” processes, while TATA-containing genes are more often highly regulated, such as by biotic or stress stimuli. These results reveal unexpected similarities in the occurrence of specific core promoter types and in their associated biological processes in yeast and humans and point to novel vertebrate-specific DNA motifs that might play a selective role in TATA-independent transcription.

[f]*Current address:* WiCell Research Institute, Madison, WI 53726

[‡]Contributed equally.

[*]Corresponding author: Ernest Martinez, Department of Biochemistry, University of California, Riverside, CA 92521, USA. Tel.: (951) 827 2031; fax: (951) 827 4434, E-mail address: ernest.martinez@ucr.edu.

Keywords

Genome-wide computational analysis; core promoter elements; Sp1; ELK-1; M22; motif distribution; transcription

1. Introduction

Accurate transcription initiation of eukaryotic genes requires a DNA region, referred to as the core promoter, which includes the transcription start site (TSS) and immediately flanking sequences. Most eukaryotic genes, including all protein-coding genes, are transcribed by RNA polymerase II (RNAPII) and are referred to as class II genes. Class II core promoters generally extend from ~40 base pairs (bp) upstream (-40) to ~40 bp downstream (+40) of the TSS (+1) and contain different combinations of various functional DNA motifs referred to as core promoter elements. These core promoter elements direct the recruitment and assembly of the class II basal/general transcription factors (TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH) and RNAPII into a functional pre-initiation complex (PIC) at the TSS and thus determine the intrinsic "basal" (i.e., unregulated) transcription activity of the core promoter (Roeder, 1998). The specific core promoter sequence also influences the transcriptional response of a given gene to particular enhancers and gene-specific transcription regulators. Thus, the information contained within the DNA sequence of the core promoter is critical for the proper regulation of gene-selective transcription in eukaryotes albeit via mechanisms that remain poorly understood (Smale and Kadonaga, 2003).

Class II core promoter elements have been best characterized in metazoan genes. They include: (i) the TATA box located at -30 relative to the TSS (+1), which is directly bound by the TATA-binding protein (TBP) subunit of the TFIID complex; (ii) the initiator (INR) element located at, or immediately adjacent to, the TSS, which is recognized by the TBP-associated factors TAF1 and TAF2 of the TFIID complex; (iii) the TFIIB recognition element (BRE) immediately flanking the TATA box and directly bound by TFIIB; and (iv) the downstream promoter element (DPE) centered at +30 downstream of the TSS, which is recognized by the TFIID subunits TAF6 and TAF9 (see Fig.1 for consensus sequences). Additional less extensively characterized core promoter sequences downstream of the TSS of specific viral and metazoan genes have been reported to influence core promoter activity and appear to be also recognized by TFIID/TAFs (Smale and Kadonaga, 2003;Lim et al., 2004;Deng and Roberts, 2005;Lewis et al., 2005;Lee et al., 2005; and references therein). Notably, none of the core promoter elements identified thus far is ubiquitous or universally required for transcription.

In yeast, core promoters still remain poorly characterized and, except for the TATA box (which in *S. cerevisiae* is generally located between -40 and -120 relative to the TSS), the other metazoan core promoter elements are generally thought to be absent. Nevertheless, specific DNA sequences have long been known to determine the position of the TSS in a small number of yeast genes and include the purine (R)-rich consensus sequence RR<u>Y</u>RR, where the underlined pyrimidine corresponds to the initiation site, and the consensus sequence T<u>CR</u>A, where either C and/or R are the initiation sites (Chen and Struhl, 1985;Hahn et al., 1985; Mosch et al., 1992; Hampsey, 1998;Smale and Kadonaga, 2003). More recently, an extended A-rich consensus sequence A(A-rich)$_5$NY<u>A</u>(A/T)NN(Arich)$_6$ has been derived from a 5'-SAGE analysis of TSSs in 2231 yeast genes (Zhang and Dietrich, 2005). Although these sequences have been referred to as yeast "initiators", their incidence in core promoters genome-wide is unclear and they are thought to function differently from metazoan INR elements (reviewed in Smale and Kadonaga, 2003). Here we use the term initiator (INR) to refer exclusively to the mammalian consensus INR sequence YYANWYY. Significantly, as in higher eukaryotes, yeast TAFs are important for core promoter-dependent transcription regulation, suggesting the existence of still unidentified TAF-dependent core promoter motifs in yeast (Green, 2000).

The binding of TBP/TFIID to the core promoter is a critical step in stable PIC assembly. Accordingly, a widely accepted model for PIC assembly at class II promoters involves the direct binding of TBP to the -30 region as an essential (and generally first) step in PIC assembly, which nucleates the recruitment of the other general/basal factors and ultimately RNAPII. In this model TBP-DNA interactions are essential for PIC assembly whether or not a TATA box is present, which is supported by the fact that TBP binds specifically and in a functional manner to a wide variety of DNA sequences that significantly diverge from the canonical TATA box sequence TATAAA (Hahn et al., 1989;Singer et al., 1990;Wobbe and Struhl, 1990;Wiley et al., 1992;Zenzie-Gregory et al., 1993;Aso et al., 1994;Kraus et al., 1996;Weis and Reinberg, 1997;Patikoglou et al., 1999). The binding of TBP to various TATA sequences induces a dramatic DNA bend (Patikoglou et al., 1999) and is stabilized by cooperative interactions with TFIIB and TFIIA, which contact flanking DNA, and with TAFs, which interact with the INR and other downstream core promoter elements (reviewed in Hahn, 2004).

The universality of this model, however, and more specifically the essential role of TBP-DNA interactions in PIC assembly, has been challenged by in vitro transcription experiments in mammalian systems indicating that the TATA-binding activity of TBP is dispensable, while TAFs are essential for basal and activated transcription from an INR-dependent “TATA-less” core promoter (Martinez et al., 1994;1995). Furthermore, basal transcription from both mammalian and *Drosophila* TATA-less core promoters requires additional cofactors distinct from TAFs/TFIID and the general transcription factors (Martinez et al., 1998;Willy et al., 2000). Thus, TATA-less core promoters that lack AT-rich sequences in the -30 region and do not stably bind TBP are likely to assemble PICs via alternative pathways and to be regulated by distinct mechanisms (Smale and Kadonaga, 2003). However, the number of such bona fide TATA-less genes remains unclear in eukaryotic genomes.

Computational analyses of 205 experimentally-defined core promoters and 1941 putative core promoter regions in *Drosophila* indicated that about 43% and 33.9%, respectively, contain the TATA box consensus sequence TATAAA or a sequence matching 5 out of the 6 consensus nucleotides, while about 67% contained the Drosophila INR element (Kutach and Kadonaga, 2000; Ohler et al., 2002). This suggested for the first time the low frequency of TATA elements and the relative abundance of INR-containing promoters in the *Drosophila* genome. Interestingly, recent analyses of *Saccharomyces* genomes also revealed that the canonical TATA box (TATAWAWR) is present in ~20% of yeast genes (Basehoar et al., 2004) suggesting that specific transcription initiation at most yeast promoters might rely on other as yet unidentified core promoter elements.

Similar studies of the frequency of the TATA box in human promoters have yielded apparently conflicting results. Analyses of ~1800 experimentally characterized promoters in the eukaryotic promoter database (EPD) indicated that 11.6% (Bajic et al., 2003), 21.8% (Gershenzon and Ioshikhes, 2005), and 76% (Trinklein et al. 2003) have a TATA box. However, the EPD database is relatively small and appears “enriched” in TATA-containing core promoters that are more amenable to experimental TSS mapping techniques (discussed in Gershenzon and Ioshikhes, 2005). Indeed, analyses of larger databases, including the database of transcription start sites (DBTSS, Suzuki et al., 2001a,2004), obtained by aligning the 5’end of full-length cDNAs to the human genome sequence, revealed a more restricted number of TATA-containing genes, although the actual percentages still varied greatly between studies - i.e., 2.6% (Fitzgerald et al., 2004), 10.4% (Gershenzon and Ioshikhes, 2005), 11-17% (Kimura et al., 2006;Jin et al., 2006), and up to 64% (Trinklein et al. 2003). Studies on the frequency of other core promoter elements in human genes also varied. One study found no preferential positioning for the consensus INR and DPE sequences in 13,010 human promoters (Fitzgerald et al., 2004), while two more recent studies found 49-63% INR-containing promoters, 22-25% BRE-containing promoters, and 12-25% DPE-containing promoters (Gershenzon and Ioshikhes, 2005;Jin et al., 2006).

Here, in an attempt to address some of the ambiguities noted above, we performed genome-scale computational analyses of human core promoters in the UCSC GoldenPath (15,685 genes) and DBTSS (10,271 genes) databases and compared the annotated biological functions of human genes with different core promoter structures. In contrast to previous analyses, we searched human core promoters both with the canonical 8-mer TATA consensus sequence (TATAWAWR) and with a list of 532 different 8-mer TATA-like elements that fit the structural definition of the TATA-TBP interface (Patikoglou et al., 1999) and most often occur at -30 in human core promoters. Our results are consistent with and extend previous observations on the frequency of TATA elements in human core promoters and further identify DNA motifs selectively enriched in TATA-less core promoters. In addition, we show that human genes with distinct core promoter structures, as defined by the presence or absence of TATA and/or INR elements, tend to control different biological processes. Unexpectedly, elements matching the metazoan consensus INR sequence also cluster specifically in the TSS region of many yeast genes suggesting that specific transcription initiation at a large fraction of eukaryotic promoters, from yeast to human, might involve similar INR elements and thus might be more conserved than previously thought.

2. Materials and Methods

The structure of a metazoan class II core promoter indicating the positioning and consensus sequences of the core promoter elements analyzed here is shown in Fig.1. Promoter regions from -2000 to +2000 relative to the TSS (+1) of all the annotated human genes were retrieved from the UCSC GoldenPath database (http://genome.ucsc.edu/downloads.html, version May, 2004). Redundant entries were randomly removed yielding 15,685 human gene promoters, which were used for the analyses of Fig. 2. Human promoter regions (-1000 to +200) were also retrieved from the DBTSS “database of transcription start sites” (ftp://ftp.hgc.jp/pub/hgc/db/dbtss/) and redundant entries were randomly removed yielding 10,271 human gene promoters, which were used in Figures 3-7. Promoter regions (-500 to +500) of *S. cerevisiae* (6165 genes) and *S. pombe* (5095 genes) were retrieved from the NCBI database (http://www.ncbi.nih.gov/Genomes/). *Drosophila* promoter regions for all genes in the NCBI database (http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=7227) were retrieved and redundant entries removed yielding 13,923 promoters. The top strand of all promoter regions were searched with the Knuth-Morris-Pratt (KMP) algorithm (Cormen et al., 1992) for consensus sequences within the windows indicated in each figure. The consensus sequences included the mammalian INR (YYANWYY), the BRE (SSRCGCC), the DPE (RGWCGTG), the canonical TATA box sequence TATAWAWR (TATA-8) (Smale and Kadonaga, 2003), and a more degenerate consensus sequence (TATA-532) which corresponds to HWHWWWWR - 576 sequences best matching the statistical and structural/functional definition of the TATA box (Bucher, 1990;Wobbe and Struhl, 1990;Patikoglou et al., 1999) - minus 44 specific sequences (see Fig.1B) that were excluded based on either functional studies (Wobbe and Struhl, 1990), their tendency to form “rigid” non-functional sequences (e.g., A-/T-tracts), similarities/overlap with INR sequences (i.e., containing CANA/T), and because they did not specifically cluster at - 30 in human core promoters (data not shown). The frequency of occurrence in the -100/+50 promoter region of the selected TATA-532 elements (divided into 4 groups according to their frequency, see supplementary Fig. S4) correlated with their specific positioning (clustering) at - 30. The most frequent groups of TATA-532 elements were also the most “specifically enriched” at -30 (data not shown). As another measure of specificity, the frequency of TATA-532 elements in the 10,271 non-redundant human core promoter regions

(-80/+80) from DBTSS was found to be significantly higher than in 10,271 “randomized” -80/+80 human promoter sequences (i.e., ~24% versus ~10.9%). Thus, the TATA-532 elements should include the most frequent/favored TATA sequences for TBP binding although they represent only a subset (~9%) of all possible 8-mer sequences that can, theoretically, expose a minor groove surface complementary/permissive to the underside of the TBP molecular saddle (see Patikoglou et al., 1999). Consensus DNA motifs highly conserved between human, mouse, rat, and dog promoters (Xie et al., 2005), were searched in human core promoters that were separated into four different categories according to the presence or absence of TATA-532 and INR elements (see Fig. 6 legend).

3. Results and Discussion

3.1 Distinct GC/AT content profiles for human TATA-containing and TATA-less promoters

Human promoters are known to have a high GC content around the TSS (Majewski and Ott, 2002;Kel-Margoulis et al., 2003;Aerts et al., 2004), which is in accord with the fact that most mammalian promoters are located within CpG islands (Gardiner-Garden and Frommer, 1987;Larsen et al., 1992;Yamashita et al., 2005;Saxonov et al., 2006). This is also illustrated here by the average GC/AT content profile of 15,685 human promoter regions in the UCSC GoldenPath database (Fig. 2A). The average GC content monotonically increases from ~45% at position -2000 to ~60% GC at +1 and then decreases back to ~45% at +2000. Within the narrower -250 to +250 region surrounding the TSS, the GC content varies between ~55% and 60% (Fig. 2B). In order to address whether the GC content profile of human core promoters differs between TATA-containing and TATA-less genes, a promoter region surrounding the TSS (i.e., -100 to +50) for each of the non-redundant 15,685 genes present in the UCSC database was searched for the presence of one of 532 different “TATA-like” sequences (see Fig. 1B and Materials and Methods). This yielded 10,608 (68%) TATA-less genes and 5077 (32%) TATA-containing genes. Analysis of the distribution of the TATA-532 elements in the UCSC database revealed, as expected, a clustering specifically at -30 (data not shown; see Fig. 3A for a profile in the DBTSS database). The GC/AT content profile within the -250 to +250 region for both groups of genes was analyzed as above. Interestingly, the GC/AT content profile of TATA-containing promoters is significantly different from that of TATA-less promoters. On average the upstream promoter region (-250 to +1) of TATA-containing genes has a relatively high AT content (~54%) which decreases to ~50% downstream of the TSS (Fig. 2C). In contrast, both the upstream and downstream promoter regions of TATA-less genes have a high GC content (~60%, Fig. 2D) and an average GC/AT distribution similar to the profile of “total” human genes (compare Fig. 2D to 2B), consistent with the above results indicating that a majority of human genes are TATA-less. Interestingly, the average GC/AT content profile of human TATA-containing promoters is similar to the overall profile of *Drosophila* promoters, which are known to be AT-rich (see supplementary Fig. S1). It is important to note that human TATA-containing promoters are AT-rich selectively upstream of the TSS and of the window used for the TATA-532 search (-100 to +50), suggesting a potential functional significance to the AT-richness. Furthermore, analysis of a select group of TATA-containing genes (interferon, interleukin, and related genes) from the experimentally-annotated EPD database confirmed the AT richness and GC/AT content asymmetry on the level of individual TATA-containing genes (supplementary Fig. S2). These results reveal for the first time an asymmetry in the GC/AT content profile of most human TATA-containing promoters, which could potentially serve as a “signature” for *in silico* recognition of this type of promoter. The functional relevance of AT-rich promoters is unknown but it is of interest to note that “rigid” DNA sequences, including poly(dA-dT) tracts incorporate poorly into nucleosomes (Kunkel and Martinson, 1981;Shimizu et al., 2000) and that a low nucleosome density is a common feature of many yeast promoters, which are AT-rich (Lee et al., 2004;Sekinger et al., 2005;Yuan et al., 2005).

3.2 Frequency profiles of core promoter elements in human and yeast promoters

To further address the frequency distribution of TATA, INR, BRE and DPE elements in human and yeast promoters, we analyzed the 10,271 non-redundant human genes in the well-annotated DBTSS database and the complete set of genes from *S. cerevisiae* and *S. pombe* in the NCBI database. Indeed, comparative distribution and clustering analyses of TATA and INR elements indicated that the annotation of human TSSs in the DBTSS database is generally more accurate than in the UCSC GoldenPath database (see supplementary Fig. S3).

We first analyzed the frequency distribution of TATA-532 elements and of the canonical TATA consensus sequence TATAWAWR (TATA-8) in human promoters in DBTSS from -250 to +150 relative to the TSS. As shown in Fig. 3A, both TATA-532 and TATA-8 elements most often occur and tightly cluster at -30, as expected. However, there are obvious differences in the frequency profiles. While TATA-8 elements are rarely found outside -30, the TATA-532 sequences occur with a frequency that increases gradually towards more distal positions and preferentially in the upstream promoter region (between -80 and -250). Furthermore, the number of TATA-532 sequences between -15 and -45 is ~10 times higher than the number of TATA-8 elements, indicating that human TATA-containing genes rarely use the canonical TATA box sequence TATAWAWR (see also supplementary Fig. S4, and below). Similar analyses of the TATA-8 consensus in yeast promoter regions revealed a significant frequency of occurrence of TATA-8 elements over an extended upstream promoter region (from -250 to +1) in both *S. pombe* and *S. cerevisiae* (Fig. 3B and C, top panel). This drastically contrasts with human promoters and correlates with the higher AT content of yeast intergenic/promoter regions (Dujon, 1996). TATA-8 sequences were found most frequently within the -20 to -150 region in *S. cerevisiae*, consistent with previous observations (Basehoar et al., 2004).

We noted two peaks of increased TATA-8 frequency in the promoter regions of both *S. pombe* (at about -45 and -135) and *S. cerevisiae* (at about -45 and -125). We reasoned that this might be due to the fact that almost half of the *S. cerevisiae* genes in the NCBI database do not have a well-defined TSS; in these genes the annotated +1 corresponds to the first nucleotide (nt) of the open reading frame (i.e., of the ATG) which might reside at a significant distance downstream of the true TSS. To address this, we separated all *S. cerevisiae* genes into those that contain an ATG at +1 and those that do not, and repeated the TATA-8 frequency profile analysis. As shown in Fig. 3C, *S. cerevisiae* promoters for which the annotated +1 is not an ATG - and thus more likely corresponds to the true TSS - only the -45 peak of TATA-8 elements was observed (middle panel) while only the -135 peak was observed in promoter regions where +1 corresponds to the ATG (bottom panel). Interestingly, in both groups of promoters the occurrence of TATA-8 elements abruptly drops at +1 and remains low in downstream sequences (+1 to +150) in both *S. cerevisiae* and *S. pombe* genes, thus revealing a sharp boundary of potential functional significance.

The frequency profile of the mammalian INR consensus element was analyzed as above in the promoter region of human, *Drosophila*, and yeast genes. As shown in Fig. 4A, we found a marked position bias for the INR at the TSS (+1) region of human promoters. *Drosophila* promoters also displayed the same position bias for the mammalian INR (Fig. 4B). Unexpectedly, sequence elements matching the mammalian INR consensus were also found most frequently in the +1 region of both *S. cerevisiae* and *S. pombe* promoters (Fig. 4C and D). Interestingly, in *S. cerevisiae* the INR had a position bias at +1 in “non-ATG start site” promoters (Fig. 4D, middle panel) and at about -3 in promoters where +1 corresponds to the ATG (bottom panel). The latter observation suggests that a significant number of *S. cerevisiae* transcripts may have a very short 5’ UTR, which is consistent with a recent study indicating that the 5’ UTR of many yeast genes (including many ribosomal protein genes) are less than 15 nt in length (David et al., 2006). This unexpected genome-wide position bias of mammalian-type INR sequences in yeast core promoters strongly suggests a possible functional conservation and contrasts with the broader distribution of the TATA box in yeast core promoters. While functional assays will be needed to directly verify the role of these INR sequences in the transcription of yeast genes, we note that the “mammalian” INR consensus sequence YYANWYY - which is conserved in *Drosophila* (Fig. 4B) - is similar to both the TCRA consensus motif and, on the opposite strand, the bi-directional RRYRR motif, which have been found to function in TSS positioning in specific yeast promoters (Hahn et al., 1985; Mosch et al., 1992). Furthermore, a mammalian-like INR element (TCACTgC) has been described previously in the core promoter of the yeast Gal80 gene and has been shown to function in a TATA-independent manner (Sakurai et al., 1994). Thus, all of these previously identified yeast “initiator” motifs (TCRA, RRYRR, TCACTgC) may simply be variations of the classical INR sequence characterized in mammals and flies, suggesting that, just like the TATA box, the sequence and function of INR elements may have been conserved to a significant degree from yeast to human. This would be consistent with the structural and functional conservation in the basal RNAPII transcription machinery, including the established core promoter-selective role of TAFs in yeast (Green, 2000). It will be of interest to determine whether, as in metazoans, yeast TAF1 and TAF2 function through the INR or whether the conservation of the INR sequence in yeast simply reflects a similar sequence preference for binding and/or initiation by RNAPII. The above results also suggest the possible existence of two major classes of yeast core promoters: those having a long 5’UTR and a short distance (~45 bp) between the TATA and the INR/TSS and those having a short 5’ UTR and a long distance (~125 bp) between the TATA box and the INR/TSS. Interestingly, while the median 5’UTR in *S. cerevisiae* promoters is 68 nt, shorter UTRs tend to be found in “housekeeping” genes while longer UTRs tend to be associated with more “regulated” genes (David et al., 2006).

Compared to the TATA and INR results, analyses of the frequency profiles of BRE and DPE consensus sequences in human and mouse promoters revealed a much lower degree of specificity in their positioning, including a high background, and a relatively low frequency of the DPE (supplementary Fig. S5). This may be due in part to the fact that the functional consensus sequences for BRE and DPE motifs are still poorly defined. We note, however, that a recent analysis of conserved human and mouse orthologous promoter sequences has estimated that only ~12% of human promoters have a DPE, while ~22% have a BRE (Jin et al., 2006). We further found that BRE and DPE sequences are largely absent in yeast promoters (data not shown).

3.3 Frequencies of different core promoter types in human and yeast genomes

Based on the above analyses of core promoters in human and yeast genes in the DBTSS and NCBI databases, we estimated the fraction of genes that have various combinations of TATA and/or INR elements within their core promoter region. Since the range of transcription start site scattering in human promoters is on average 62 ± 20bp (Suzuki et al, 2001b), a window from -80 to +80 was selected to search for human genes that have at least one TATA-like sequence (i.e., one of the TATA-532 elements) but no INR (“TATA only” category); an INR consensus but no TATA-like sequence (“INR only”); both TATA and INR elements (“TATA +INR”); and none of these elements (“None” category). As summarized in Fig. 5A, of 10,271 human genes in DBTSS we found that almost half - i.e., 4730 genes (46% of total) - have neither a TATA-532 sequence nor an INR element and almost one third - i.e., 3107 genes (30%) - have only INR elements. Thus, the vast majority (~76%) of human core promoters lack TATA-like sequences, consistent with the results obtained with the UCSC database (see Section 3.1). Furthermore, only 865 genes (8%) belong to the “TATA only” category and 1569 genes (15%) have both a TATA-like sequence and an INR element at any position relative to each other. Surprisingly, only 211 genes (2%) have TATA and INR elements that are positioned relative to each other in a “synergistic configuration” (O’Shea-Greenfield and Smale, 1992) -

i.e., the INR positioned 15-30 bp downstream of the TATA sequence (sub-category not illustrated in Fig. 5A). Thus, many promoters that have both TATA and INR elements within their -80/+80 region might have several TSSs and core promoter elements functioning independently or, alternatively, some of the TATA and INR elements might not be functional and these promoters/genes might belong to another category.

We found that about half of human promoters/genes in DBTSS (~46%) have the 7-mer INR consensus element within their -80/+80 region. This is likely to be a conservative estimate of INR-containing promoters since certain deviations from the perfect consensus INR are known to be functional (Javahery et al., 1994) and ~49% of promoters in DBTSS have been shown to have an 8-mer INR element as defined by a PWM in a much more restricted TSS region (Gershenzon and Ioshikhes, 2005). Furthermore, our results indicating that only ~24% of human core promoters have TATA-like sequences (~10% if one considers only the -45/-15 region) - with only ~10% of these being canonical TATA-8 elements (TATAWAWR) - are consistent with previous observations indicating that a very small fraction (~2.6%) of human promoters have the TATA consensus sequence TATAAAD (Fitzgerald et al., 2004), while ~10-17% have more degenerate TATA elements as defined by a PWM for the TATA box (Bucher, 1990;Gershenzon and Ioshikhes, 2005;Jin et al., 2006;Kimura et al., 2006). Moreover, our observation that both the overall AT content of human promoters (Fig. 2A) and the occurrence of TATA-532 elements significantly increase upstream of -80 (Fig. 3A), might explain the high percentage (~64%) of TATA-containing promoters reported by a previous study that searched for TAWWWW sequences in a relatively broad (-550/+50) window (Trinklein et al., 2003). It is also important to note that we have obtained comparable results with two independent databases (DBTSS and UCSC, see above and section 3.1) and therefore the scarcity of human TATA-containing core promoters reported here is not a peculiarity of the annotated database used.

A similar analysis was performed on the 6165 genes of yeast *S. cerevisiae* in the NCBI database. TATA sequences and INR elements were searched within the -150/+1 and -25/+25 promoter regions, respectively. As expected from the high AT content of yeast promoters, virtually all (more than 96%) yeast core promoters were found to have at least one of the TATA-532 elements between -150 and +1 (data not shown). However, only ~24% of yeast promoters have the canonical TATAWAWR (TATA-8) elements (Fig. 5B), which is consistent with a previous estimate (Basehoar et al., 2004). In contrast, ~40% of yeast core promoters have mammalian-type INR sequences within the -25/+25 TSS region (Fig. 5B), which is similar to the fraction (~46%) of human core promoters containing INR elements. This similarity is unexpected considering the marked differences in overall GC content between human and yeast promoter regions (50-60% vs. 30-40% GC, respectively). Furthermore, considering the recent evidence suggesting that yeast genes lacking canonical TATAWAWR elements are functionally “TATA-less” in vivo (Basehoar et al., 2004), an intriguing conservation in the proportion of the different core promoter types in yeast and human genomes becomes apparent (Fig. 5A and B). Altogether, these results suggest the possibility that transcription of most eukaryotic genes from yeast to human could potentially occur via TATA-independent pathways and many may utilize a conserved INR element. These results also suggest that additional core promoter motifs may play a role in specific transcription initiation at a significant number (~46%) of human and yeast promoters that lack both TATA and INR elements.

3.4 DNA sequence motifs that occur preferentially in TATA-less core promoters

A recent study of mammalian promoter regions identified several DNA motifs that are highly conserved across the human, mouse, rat, and dog genomes (Xie et al., 2005). In order to determine whether any of these elements could play a role in directing transcription from

TATA-less promoters, we searched the -250/+150 region of the four categories of human promoters described in Fig.5A - i.e., “TATA only”, “TATA+INR”, “INR only”, and “None” categories - with the 11 most highly conserved mammalian consensus motifs (M1 to M11) and six additional motifs (M14, M16, M21, M22, M30 and M48) that tend to occur preferentially within the core promoter region (Xie et al., 2005). The number and percentage of genes in each of the four categories that have at least one of the above 17 motifs within the -250/+150 region are presented in supplementary Fig. S6. Three motifs were enriched in TATA-less promoters: the M3 motif corresponding to a consensus ELK-1 binding site (SCGGAAGY), the M6 motif corresponding to an Sp1 binding motif (GGGCGGR), and M22, a novel motif (TGCGCANK). Moreover, as shown in Fig. 6A, 60.5% of promoters containing M3 (ELK-1), 62.8% of promoters containing M6 (Sp1), and 70.9% of promoters containing M22 belong to the “None” category of promoters. These percentages are all significantly higher than the fraction (46.1%) of total promoters that belong to the “None” category in the human genome ($P < 0.05$). This indicates that promoters containing these motifs within the -250/+150 region tend to lack both TATA and INR elements within their -80/+80 core promoter region. Furthermore, this core promoter bias is significantly more pronounced for M22 than for M3 (ELK-1) and M6 (Sp1) motifs ($P < 0.05$).

Interestingly, only 19.4% of M22-containing promoters belong to the “INR only” category. This percentage is significantly lower ($P < 0.05$) than the 26.3% and 29.6% of M6 (Sp1)- and M3 (ELK-1)-containing promoters in this category, respectively, and than the overall occurrence of “INR only” core promoters (30.3%) in the human genome (Fig. 6A, column INR only). This suggests possible redundant roles of M22 and INR elements at TATA-less promoters. To further address this, we analyzed the frequency distribution profile of M22 and the other two motifs in TATA-containing and TATA-less promoters. We found that M3 (ELK-1) and M6 (Sp1) have a position bias at about -20 and -50, respectively, in both TATA-containing and TATA-less promoters (Fig. 6B and C), in agreement with their global position bias in the human genome reported by Xie et al. (2005) (-24 and -63, respectively; see supplementary Fig. S6). In contrast, M22 was found more frequently in the +1 region of TATA-less promoters. This preferred positioning of M22 was not observed in TATA-containing promoters (Fig. 6D) and slightly differs from the global position bias of M22 (-17) in the whole human genome (Xie et al., 2005). Notably, the occurrence of M22 in TATA-less promoters decreases sharply downstream of +1, which could reflect a possible function of M22 motifs at the TSS and would be consistent with the reduced frequency of initiator elements in M22-containing TATA-less core promoters. In contrast to the sharp drop off in M22 frequency downstream of +1, there was a more graded decrease in M22 elements in the upstream region, which could reflect a prevalence of multiple TSSs among M22-containing TATA-less promoters. Finally, all three elements enriched in human TATA-less promoters -M3 (ELK-1), M6 (Sp1) and M22 - are GC-rich and are not found in the promoters of *Drosophila* or yeast, which are AT-rich (data not shown).

3.5 Gene Ontology of human genes with different core promoter structures

To determine whether the presence or absence of TATA and/or INR elements correlates with specific biological functions of the cognate genes, we compared the functions of human genes in the different core promoter categories using Gene Ontology (GO) (Figure 7). Using high confidence cut-off values (EASE scores < 0.0001) we found that the most over-represented genes in each promoter category were associated with specific biological processes that differed between the distinct categories (Fig. 7A; see supplementary Fig. S7 for the complete list of genes). For example, the “TATA only” category had a selective enrichment of genes involved in organogenesis and response to biotic stimuli (including immune- and stress-response genes), while the “TATA+INR” category of genes having the TATA box 15 to 30 nucleotides upstream of the INR (TATA+INR module) had an over-representation of genes involved in nucleosome assembly (e.g., histone genes) and cell adhesion. In contrast, the “INR only” category was predominantly enriched for genes involved in basic biological processes such as protein biosynthesis and mRNA processing, while the “None” category was enriched for genes involved in other basic processes such as intracellular transport, cell growth and maintenance and protein metabolism. The notion of “TATA only” genes differing from genes in other core promoter categories was reinforced when a larger number of Biological Processes (EASE score < 0.001) was compared between the “TATA only” and the “None” categories. Fig. 6B shows that there were only a few biological processes, such as cell proliferation and cell cycle, in which the EASE scores of the “TATA only” category were similar to those of the “None” category. In contrast, there was more overlap in EASE scores between the “INR only” and the “None” categories (supplementary Fig. S8). Furthermore, when the GO cellular component was analyzed “TATA only” and “None” categories of genes were most often associated with the extracellular and intracellular compartments, respectively (supplementary Fig. S9). Overall, these results indicate that genes containing different core promoter elements tend to control different biological processes and are consistent with the general notion that promoters of housekeeping genes in vertebrate organisms are often TATA-less and/or associated with CpG islands, while cell type-specific or highly regulated genes often have TATA boxes (Gardiner-Garden and Frommer, 1987;Larsen et al., 1992;Smale and Kadonaga, 2003;Yamashita et al., 2005).

Kimura et al. (2006) recently estimated that about 52% of human genes have putative alternative promoters, i.e., clusters of start sites separated by 500 bp or more. For those genes only one annotated promoter was randomly selected for our data set. Thus, while the percentage of promoters containing TATA and/or INR elements should not be affected by considering randomly only one promoter per gene - which is consistent with the similar percentage of TATA-containing promoters in genes with one or several putative alternative promoters (Kimura et al., 2006) - it was however unexpected to find such clear differences in our Gene Ontology (GO) analyses. This suggests that a significant number of genes for some of the GO functional categories found associated with a specific core promoter type either do not have alternative promoters or have alternative promoters of the same type. Consistent with the former possibility, genes associated with the GO terms *extracellular region, G-protein coupled receptor signaling pathway, structural components of the ribosome*, and *mitochondrion* tend to lack alternative promoters (Kimura et al., 2006). This correlates very well with our results in Fig. 7 and in Fig. S9 indicating that the clearest distinctions in GO biological terms between TATA-containing and TATA-less groups correspond to the GO terms *extracellular region* and *response to external/biotic stimulus* (TATA category) and *protein biosynthesis, large ribosomal subunit*, and *mitochondrion* (TATA-less category). Interestingly, these results for the human genome are also consistent with recent data on *S. cerevisiae* genes indicating that TATA-containing genes are generally highly regulated and associated with inducible stress-related responses, while TATA-less genes have mostly housekeeping functions (Basehoar et al., 2004).

3.6 Concluding remarks

The results presented here are consistent with and complement previous observations regarding the frequency of TATA elements in human core promoters, clearly indicating that the vast majority (~76%) of human core promoter regions (from -80 to +80 relative to the TSS) not only lack canonical TATA elements described by the consensus TATAWAWR but also any of the TATA-like sequences in our TATA-532 list (Fig.1B) which is comprised of elements that fit the structural definition of the TBP-TATA interaction (Patikoglou et al., 1999) and most often occur at -30 in human TATA-containing core promoters (Fig. 3A). While a complete functionally validated list of TBP-binding sites is not available and our TATA-532 sequences certainly do not represent the totality of sites recognized by TBP, it is worth noting that our exact match search with this defined set of TATA-532 elements yields comparable results to PWM approaches used recently to analyze the DBTSS database (Gershenzon and Ioshikhes, 2005;Kimura et al., 2006) and a set of conserved orthologous human-mouse promoter pairs (Jin et al., 2006). Indeed, we find that ~10% of human core promoters in DBTSS have a TATA-like element in the restricted -45/-15 region corresponding to the peak of highest TATA-532 frequency (Fig. 3A), which is comparable to the previous estimates of ~10% (Gershenzon and Ioshikhes, 2005), 11-17% (Kimura et al., 2006), and ~17% (Jin et al., 2006) obtained by searching similar restricted -30 regions. However, because TSS annotation is not perfect in DBTSS and since many human promoters have several TSSs (perhaps driven in part by independent core promoter elements) with an average scattering of ~62 ±20 nt (Suzuki et al., 2001b), the random selection of only one TSS per gene and the analysis of narrow -30 windows could overlook a significant number of functional TATA-like elements. Thus, we consider it likely that the actual fraction of TATA-containing promoters in the human genome is larger, and based on our analysis of the -80/+80 core promoter window we estimate it to be in the range of ~24%.

It is important to note that in addition to the sequence of the TATA element, other parameters affect TBP-DNA interaction in vivo, such as the assembly of DNA into chromatin and cooperative or negative effects of DNA-binding proteins and TBP-binding factors. Indeed, TATA sequences are selectively used by TBP in vivo in a manner that does not solely depend on whether TBP can bind a particular sequence in vitro; for instance, only ~20% of yeast AT-rich promoters appear to depend on a functional TBP DNA-binding surface in vivo (Basehoar et al., 2004). Thus, while our results, in conjunction with previous studies, suggest the distinct possibility that TATA-dependent pathways for core promoter recognition might be involved at only a minority of eukaryotic genes, ultimately a more detailed understanding of genome-wide functional TBP-DNA interactions in vivo will be needed to demonstrate this.

Interestingly, we found that a similar fraction of human and yeast core promoters (~46% and ~40%, respectively) have “mammalian-type” INR consensus elements in the TSS region. Although experiments on a limited number of yeast core promoters support the possible function of such sequences (Hahn et al., 1985;Sakurai et al., 1994; Kuehner and Brow, 2006), additional experiments are required to verify their function as bona fide INR elements in the large fraction of genes identified here and to characterize the mechanisms involved. Our data nevertheless suggest the intriguing possibility that, just like the TATA box, the INR element may have been highly conserved during evolution and might control transcription of a significant fraction of genes in most eukaryotes from yeast to human. Consistent with this, are the similarities between human and yeast genomes not only in the fraction of genes in the different core promoter categories but also in the type of biological process associated with genes in these categories. Indeed, both human and yeast TATA-less genes are most often involved in basic “housekeeping” processes, while TATA-dependent genes tend to be highly regulated and responsive to stress and extracellular signals (including developmental stimuli in humans). This conservation is intriguing considering a recent report indicating that core promoter types may switch class at a significant rate between different yeast (*Saccharomyces*) species - i.e, from TATA-containing to TATA-less and vice versa (Bazykin and Kondrashov, 2006). The observed conservation described above might thus suggest that a significant number of yeast genes might not switch core promoter class as frequently as others, perhaps due to selective pressure. Consistent with this possibility, genes with promoters that switch class between *Saccharomyces* species appear to differ from other genes in the same class by either having a lower level of expression (TATA-less genes) or being less sensitive to TBP mutations (TATA-containing genes) (Bazykin and Kondrashov, 2006).

Although our results suggest that INR elements may contribute to TATA-independent transcription from ~30% of human promoters, the largest category of human core promoters

(~46%) appear to lack both TATA and INR sequences. This raises the question as to which DNA elements recruit and/or position the basal machinery on these promoters. It has long been known that Sp1 binding sites (i.e., GC-boxes) are often found in TATA-less promoters within CpG islands and that Sp1 can direct weak transcription initiation from heterogeneous TSSs in vitro from core promoters that lack both TATA and INR elements (Smale and Kadonaga, 2003). Consistent with this, we found that TATA-less promoters that lack INR elements are generally enriched in the M6 (Sp1) motif. Our results also suggest that two additional motifs, M3 (ELK-1) and M22 are also preferentially found in promoters that lack TATA and INR elements. The M3 motif corresponds to a binding site for transcription regulators of the ETS-domain family of proteins such as ELK-1. In this regard, it is of interest to note that the N-terminal DNA-binding domain of the INR-binding protein IBP39 of *Trichomonas vaginalis* adopts a general scaffold structure strikingly similar to that of metazoan ETS-family proteins that is important for TATA-less promoter function in this ancient eukaryote (Schumacher et al., 2003). The M22 motif is the most intriguing because it is preferentially located in the TSS/+1 region of TATA-less promoters and is rarely found in TATA-less promoters that contain INR elements. The potential role of M22 motifs in regulating TATA-independent transcription and the identity of M22-binding factors will be worth pursuing.

In summary, our results suggest a prevalence of INR-containing over TATA-containing promoters in the human genome, an unexpected conservation of metazoan-type INR consensus elements in the TSS region of a significant number of yeast promoters, conserved differences in the biological processes associated with TATA-less and TATA-containing genes in yeast and humans, and have pointed to novel vertebrate-specific DNA motifs that may play a core promoter-selective role at human TATA-less promoters.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank J. Evans for editorial help. C.Y. and E.B were supported in part by the GGB Graduate Program at UCR. This work was supported by grants from NSF (CCR-0309902), NSF China (60528001) and China National Key Project for Basic Research (973) (2002CB512801) to T. J. by a grant from NIH (R01DK053892) to F.M.S. and by a CAREER NSF grant (MCB-0448488) to E.M.

Abbreviations

bp, base pair; BRE, TFIIB recognition element; DPE, downstream promoter element; INR, initiator; nt, nucleotide; PIC, pre-initiation complex; PWM, position weight matrix; RNAPII, RNA polymerase II; TAF, TBP-associated factor; TBP, TATA-binding protein; TFIIB, transcription factor IIB; TSS, transcription start site.

References

Aerts S, Thijs G, Dabrowski M, Moreau Y, De Moor B. Comprehensive analysis of the base composition around the transcription start site in Metazoa. BMC Genomics 2004;5:34–44. [PubMed: 15171795]

Aso T, Conaway JW, Conaway RC. Role of core promoter structure in assembly of the RNA polymerase II preinitiation complex. A common pathway for formation of preinitiation intermediates at many TATA and TATA-less promoters. J. Biol. Chem 1994;269:26575–26583. [PubMed: 7929383]

Bajic VB, Choudhary V, Hock CK. Content analysis of the core promoter region of human genes. In Silico Biol 2004;4:109–125. [PubMed: 15089758]

Basehoar AD, Zanton SJ, Pugh BF. Identification and distinct regulation of yeast TATA box-containing genes. Cell 2004;116:699–709. [PubMed: 15006352]

Bazykin GA, Kondrashov AS. Rate of promoter class turn-over in yeast evolution. BMC Evol. Biol 2006;6:14. [PubMed: 16472383]

Bucher P. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. J. Mol. Biol 1990;212:563–578. [PubMed: 2329577]

Chen W, Struhl K. Yeast mRNA initiation sites are determined primarily by specific sequences, not by the distance from the TATA element. EMBO J 1985;4:3273–3280. [PubMed: 3912167]

Cormen, T.; Leiserson, C.; Riverst, R. Introduction to Algorithms. MIT Press; Cambridge: 1992.

David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, Jones T, Davis RW, Steinmetz LM. A high-resolution map of transcription in the yeast genome. Proc. Natl. Acad. Sci. U. S. A 2006;103:5320–5325. [PubMed: 16569694]

Deng W, Roberts SGE. A core promoter element downstream of the TATA box that is recognized by TFIIB. Genes Dev 2005;19:2418–2423. [PubMed: 16230532]

Dujon B. The yeast genome project: what did we learn? Trends Genet 1996;12:263–270. [PubMed: 8763498]

Fitzgerald PC, Shlyakhtenko A, Mir AA, Vinson C. Clustering of DNA sequences in human promoters. Genome Res 2004;14:1562–1574. [PubMed: 15256515]

Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. J. Mol. Biol 1987;196:261–282. [PubMed: 3656447]

Gershenzon NI, Ioshikhes IP. Synergy of human Pol II core promoter elements revealed by statistical sequence analysis. Bioinformatics 2005;21:1295–1300. [PubMed: 15572469]

Green MR. TBP-associated factors (TAFIIs): multiple, selective transcriptional mediators in common complexes. Trends Biochem. Sci 2000;25:59–63. [PubMed: 10664584]

Hahn S. Structure and mechanism of the RNA polymerase II transcription machinery. Nat. Struct. Mol. Biol 2004;11:394–403. [PubMed: 15114340]

Hahn S, Hoar ET, Guarente L. Each of three “TATA elements” specifies a subset of the transcription initiation sites at the CYC-1 promoter of *Saccharomyces cerevisiae*. Proc. Natl. Acad. Sci. U. S. A 1985;82:8562–8566. [PubMed: 3001709]

Hahn S, Buratowski S, Sharp PA, Guarente L. Yeast TATA-binding protein TFIID binds to TATA elements with both consensus and nonconsensus DNA sequences. Proc. Natl. Acad. Sci. U. S. A 1989;86:5718–5722. [PubMed: 2569738]

Hampsey M. Molecular genetics of the RNA polymerase II general transcriptional machinery. Microbiol. Mol. Biol. Rev 1998;62:465–503. [PubMed: 9618449]

Hosack DA, Dennis G, Sherman BT, Lane HC, Lempicki RA. Identifying biological themes within lists of genes with EASE. Genome Biology 2003;4:R60.

Javahery R, Khachi A, Lo K, Zenzie-Gregory B, Smale ST. DNA sequence requirements for transcriptional initiator activity in mammalian cells. Mol. Cell. Biol 1994;14:116–127. [PubMed: 8264580]

Jin VX, Singer GA, Agosto-Perez FJ, Liyanarachchi S, Davuluri RV. Genome-wide analysis of core promoter elements from conserved human and mouse orthologous pairs. BMC Bioinformatics 2006;7:114. [PubMed: 16522199]

Kel-Margoulis OV, Tchekmenev D, Kel AE, Goessling E, Hornischer K, Lewicki-Potapov B, Wingender E. Composition-sensitive analysis of the human genome for regulatory signals. In Silico Biol 2003;3:145–171. [PubMed: 12954097]

Kimura K, et al. Diversification of transcriptional modulation: Large-scale identification and characterization of putative alternative promoters of human genes. Genome Res 2006;16:55–65. [PubMed: 16344560]

Kraus RJ, Murray EE, Wiley SR, Zink NM, Loritz K, Gelembiuk GW, Mertz JE. Experimentally determined weight matrix definitions of the initiator and TBP binding site elements of promoters. Nucleic Acids Res 1996;24:1531–1539. [PubMed: 8628688]

Kunkel GR, Martinson HG. Nucleosomes will not form on double-stranded RNA or over poly(dA).poly (dT) tracts in recombinant DNA. Nucleic Acids Res 1981;9:6869–6888. [PubMed: 7335494]

Larsen F, Gundersen G, Lopez R, Prydz H. CpG islands as gene markers in the human genome. Genomics 1992;13:1095–1107. [PubMed: 1505946]

Lee CK, Shibata Y, Rao B, Strahl BD, Lieb JD. Evidence for nucleosome depletion at active regulatory regions genome-wide. Nat. Genet 2004;36:900–905. [PubMed: 15247917]

Lee DH, Gershenzon N, Gupta M, Ioshikhes IP, Reinberg D, Lewis BA. Functional characterization of core promoter elements: the downstream core element is recognized by TAF1. Mol. Cell. Biol 2005;25:9674–9686. [PubMed: 16227614]

Lewis BA, Sims RJ 3rd, Lane WS, Reinberg D. Functional characterization of core promoter elements: DPE-specific transcription requires the protein kinase CK2 and the PC4 coactivator. Mol. Cell 2005;18:471–481. [PubMed: 15893730]

Lim CY, Santoso B, Boulay T, Dong E, Ohler U, Kadonaga JT. The MTE, a new core promoter element for transcription by RNA polymerase II. Genes Dev 2004;18:1606–1617. [PubMed: 15231738]

Majewski J, Ott J. Distribution and characterization of regulatory elements in the human genome. Genome Res 2002;12:1827–1836. [PubMed: 12466286]

Martinez E, Chiang CM, Ge H, Roeder RG. TATA-binding protein-associated factor(s) in TFIID function through the initiator to direct basal transcription from a TATA-less class II promoter. EMBO J 1994;13:3115–3126. [PubMed: 7518774]

Martinez E, Zhou Q, L'Etoile ND, Oelgeschlager T, Berk AJ, Roeder RG. Core promoter-specific function of a mutant transcription factor TFIID defective in TATA-box binding. Proc. Natl. Acad. Sci. U. S. A 1995;92:11864–11868. [PubMed: 8524864]

Martinez E, Ge H, Tao Y, Yuan CX, Palhan V, Roeder RG. Novel cofactors and TFIIA mediate functional core promoter selectivity by the human TAFII150-containing TFIID complex. Mol. Cell. Biol 1998;18:6571–6583. [PubMed: 9774672]

O'Shea-Greenfield A, Smale ST. Roles of TATA and initiator elements in determining the start site location and direction of RNA polymerase II transcription. J. Biol. Chem 1992;267:1391–1402. [PubMed: 1730658]

Patikoglou GA, Kim JL, Sun L, Yang SH, Kodadek T, Burley SK. TATA element recognition by the TATA box-binding protein has been conserved throughout evolution. Genes Dev 1999;13:3217–3230. [PubMed: 10617571]

Roeder RG. Role of general and gene-specific cofactors in the regulation of eukaryotic transcription. Cold Spring Harb. Symp. Quant. Biol 1998;63:201–218. [PubMed: 10384284]

Sakurai H, Ohishi T, Fukasawa T. Two alternative pathways of transcription initiation in the yeast negative regulatory gene GAL80. Mol. Cell. Biol 1994;14:6819–6828. [PubMed: 7935399]

Saxonov S, Berg P, Brutlag DL. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc. Natl. Acad. Sci. U.S.A 2006;103:1412–1417. [PubMed: 16432200]

Sekinger EA, Moqtaderi Z, Struhl K. Intrinsic histone-DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast. Mol. Cell 2005;18:735–748. [PubMed: 15949447]

Shimizu M, Mori T, Sakurai T. Destabilization of nucleosomes by an unusual DNA conformation adopted by poly(dA).poly(dT) tracts in vivo. EMBO J 2000;19:3358–3365. [PubMed: 10880448]

Smale ST, Kadonaga JT. The RNA polymerase II core promoter. Annu. Rev. Biochem 2003;72:449–479. [PubMed: 12651739]

Singer VL, Wobbe CR, Struhl K. A wide variety of DNA sequences can functionally replace a yeast TATA element for transcriptional activation. Genes Dev 1990;4:636–645. [PubMed: 2163345]

Suzuki Y, Taira H, Tsunoda T, Mizushima-Sugano J, Sese J, Hata H, Ota T, Isogai T, Tanaka T, Morishita S, Okubo K, Sakaki Y, Nakamura Y, Suyama A, Sugano S. Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. EMBO Rep 2001b;2:388–393. [PubMed: 11375929]

Suzuki Y, Tsunoda T, Sese J, Taira H, Mizushima-Sugano J, Hata H, Ota T, Isogai T, Tanaka T, Nakamura Y, Suyama A, Sakaki Y, Morishita S, Okubo K, Sugano S. Identification and characterization of the potential promoter regions of 1031 kinds of human genes. Genome Res 2001a;11:677–684. [PubMed: 11337467]

Suzuki Y, Yamashita R, Sugano S, Nakai K. DBTSS, DataBase of Transcriptional Start Sites: progress report 2004. Nucleic Acids Res 2004;32:D78–81. [PubMed: 14681363]

Trinklein ND, Aldred SJ, Saldanha AJ, Myers RM. Identification and functional analysis of human transcriptional promoters. Genome Res 2003;13:308–312. [PubMed: 12566409]

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 1997;17:2973–2984. [PubMed: 9154795]

Wiley SR, Kraus RJ, Mertz JE. Functional binding of the "TATA" box binding component of transcription factor TFIID to the -30 region of TATA-less promoters. Proc. Natl. Acad. Sci. U. S. A 1992;89:5814–5818. [PubMed: 1321424]

Willy PJ, Kobayashi R, Kadonaga JT. A basal transcription factor that activates or represses transcription. Science 2000;290:982–985. [PubMed: 11062130]

Wobbe CR, Struhl K. Yeast and human TATA-binding proteins have nearly identical DNA sequence requirements for transcription in vitro. Mol. Cell. Biol 1990;10:3859–3867. [PubMed: 2196437]

Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M. Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. Nature 2005;434:338–345. [PubMed: 15735639]

Yamashita R, Suzuki Y, Sugano S, Nakai K. Genome-wide analysis reveals strong correlation between CpG islands with nearby transcription start sites of genes and their tissue specificity. Gene 2005;350:129–136. [PubMed: 15784181]

Yuan GC, Liu YJ, Dion MF, Slack MD, Wu LF, Altschuler SJ, Rando OJ. Genome-scale identification of nucleosome positions in *S. cerevisiae*. Science 2005;309:626–630. [PubMed: 15961632]

Zenzie-Gregory B, Khachi A, Garraway IP, Smale ST. Mechanism of initiator-mediated transcription: evidence for a functional interaction between the TATA-binding protein and DNA in the absence of a specific recognition sequence. Mol. Cell. Biol 1993;13:3841–3849. [PubMed: 8321191]

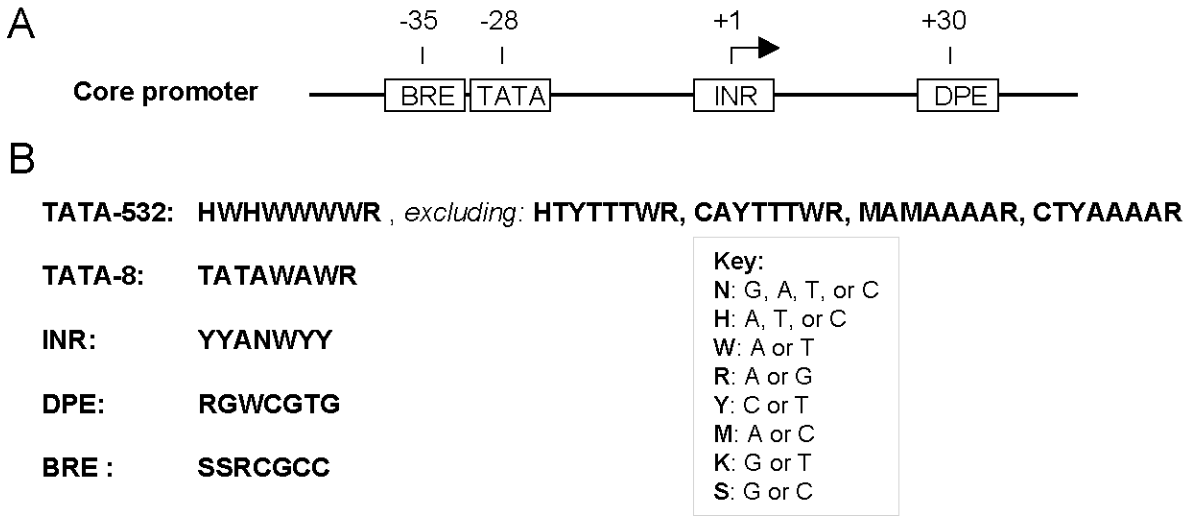

Fig 1.

Structure of the core promoter in eukaryotic genes and sequence of core promoter elements. (A) The positions in nucleotides (nt) relative to the transcription start site (TSS, +1) are given for core promoter elements: BRE, TFIIB response element; TATA, TATA box; INR, initiator element; DPE, downstream promoter element. (B) Consensus sequences of the core promoter elements used in this study. Key, IUPAC nomenclature.

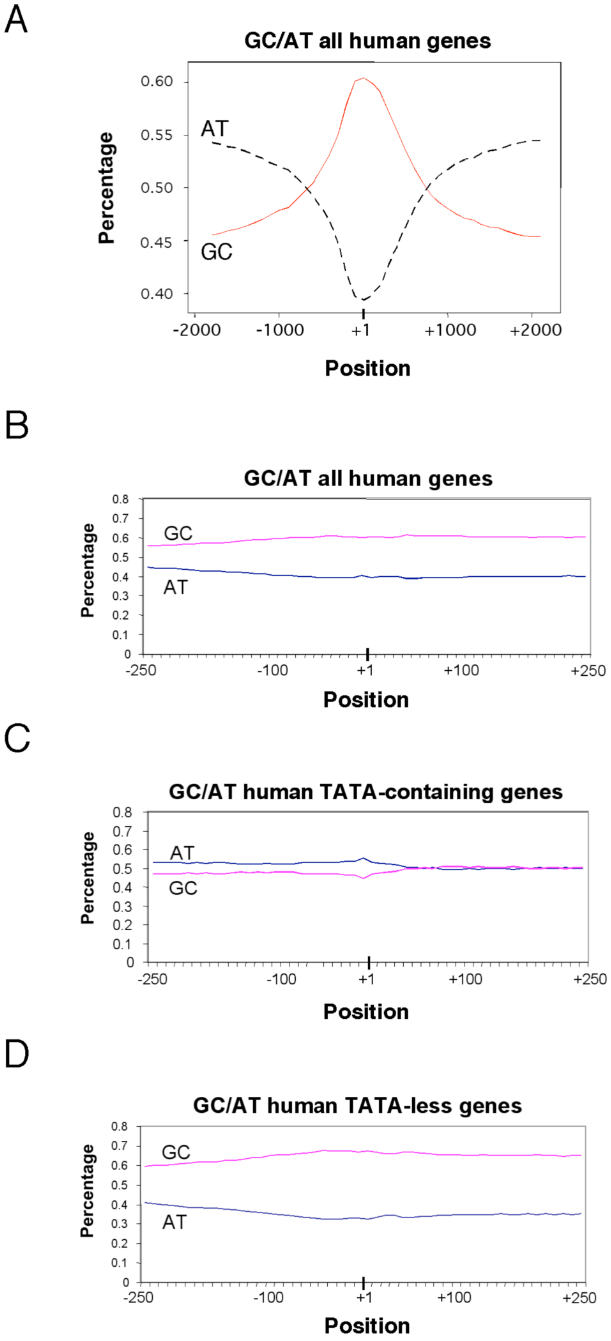

Fig 2.

GC/AT content of human promoters. (A) Average GC/AT content profile of 15,685 non-redundant human genes in the UCSC database between -2000 and +2000 relative to the TSS (+1). (B) GC/AT content profile as in (A) for the narrower region -250 to +250. (C) GC/AT content profiles as in (B) but for only those human genes in the UCSC database that contain at least one TATA-532 element between -150 and +50 (5077 genes, ~32%) as defined in Fig. 1 and the text. (D) As in (C) but for the remaining TATA-less genes (10,608 genes, ~68%).

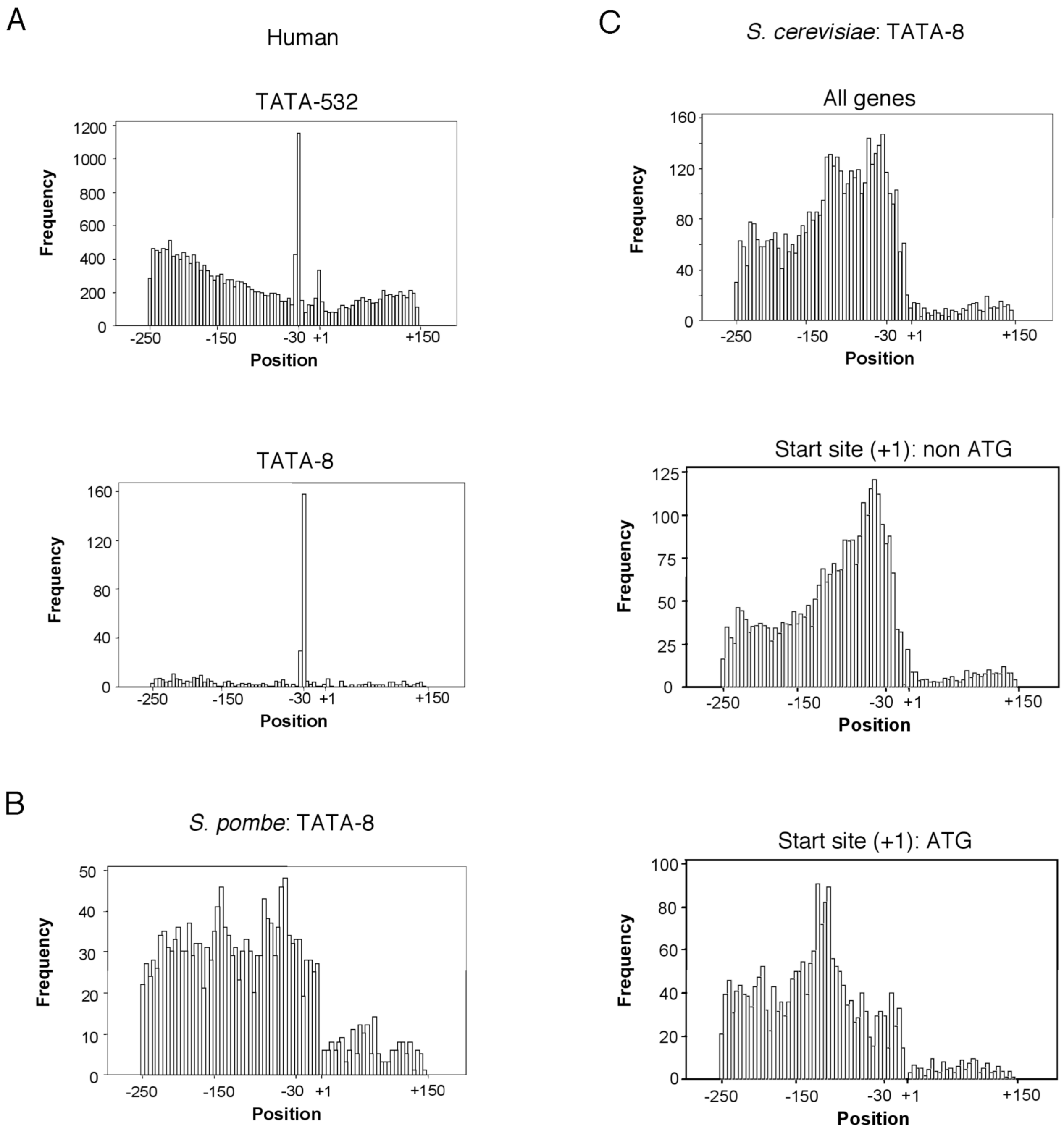

Fig 3.

Distribution of the TATA box in human and yeast promoters. (A) Frequency profile of TATA-532 elements (top) and the canonical TATA-8 consensus (bottom) within the region -250 to +150 relative to the TSS (+1) of 10,271 non redundant human genes from the DBTSS database. (B) Frequency profile of the canonical TATA-8 consensus in the -250 to +150 region of all genes in *S. pombe* (5,095 genes) from the NCBI database. (C) Frequency profile of the canonical TATA-8 consensus in the -250 to +150 region of all genes in *S. cerevisiae* from the NCBI database (6165 genes, top), as well as those *S. cerevisiae* genes that do not contain an ATG at +1 (3195 genes, middle) and those that do (2970 genes, bottom) (see text for details). Only the TATA-8 sequences were used for the profile analysis; the TATA-532 search yielded no peak. Note the difference in frequency scale on the y-axis between *S. cerevisiae* and *S. pombe*. Bin size is 5 nt.

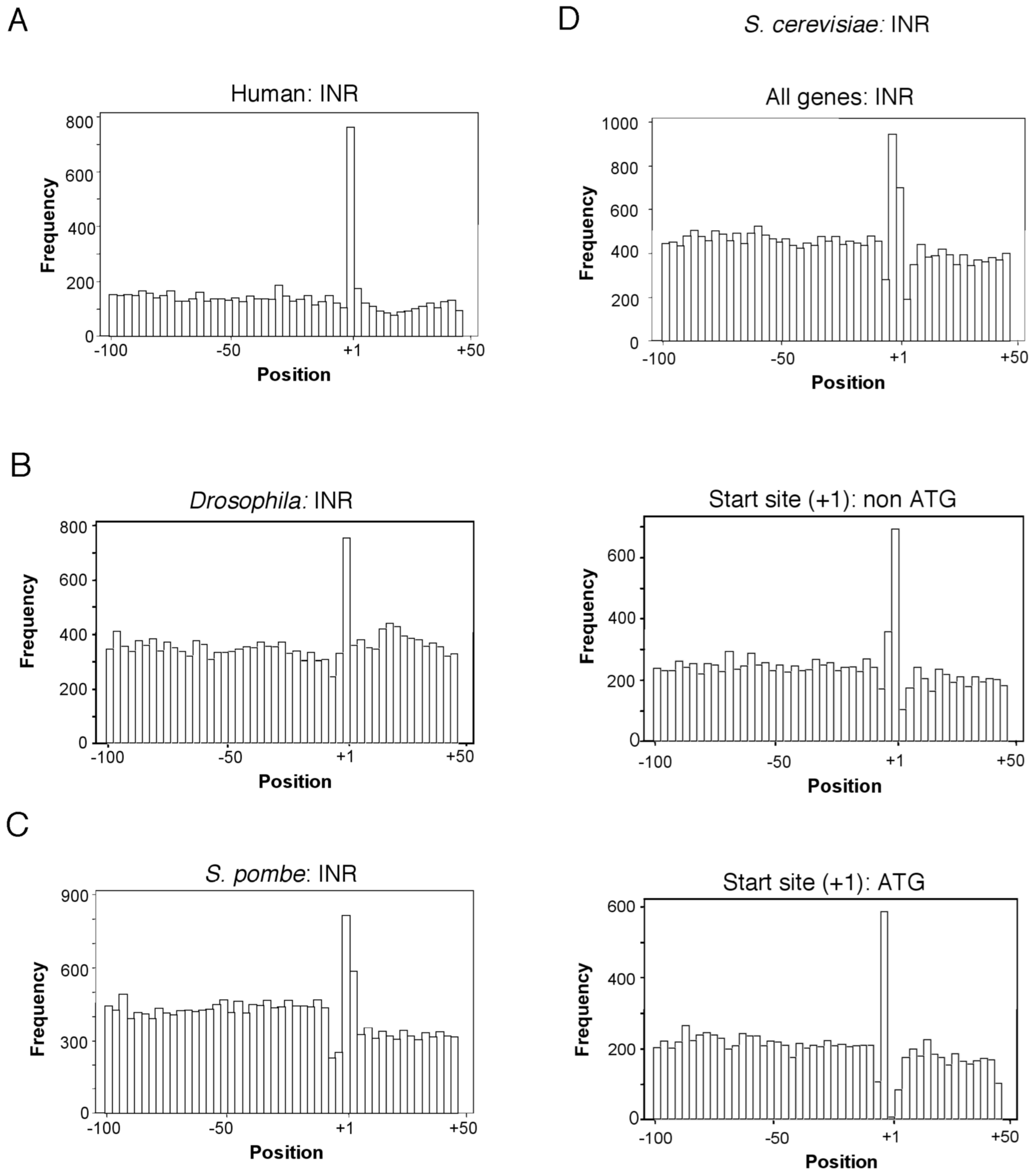

Fig 4.

Distribution of INR elements in human, Drosophila and yeast promoters. Frequency profiles of the mammalian INR (see Fig. 1) in the -100 to +50 region relative to the TSS (+1) of (A) 10,271 human genes from the DBTSS database, (B) 13,923 Drosophila genes from the NCBI database, (C) 5095 *S. pombe* genes from the NCBI database and (D) *S. cerevisiae* genes as defined in Fig. 3C. Bin size is 3 nt.

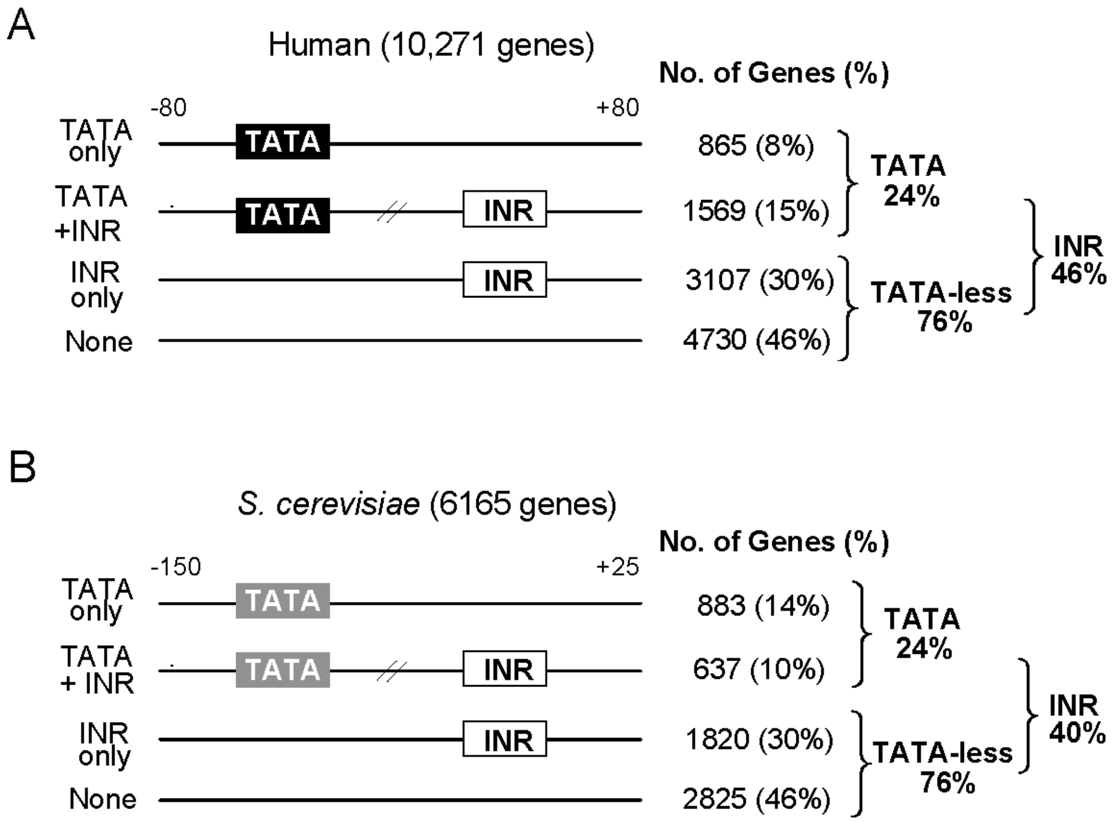

Fig 5.

Frequencies of the different categories of core promoters in human and yeast genes. (A) Human promoters (10,271 total) from DBTSS were searched by scanning a 110 nt window within the -80 to +80 region relative to TSS (+1) for the existence of TATA-532 and INR elements (see Fig. 1). TATA only, genes with a TATA box but no INR in the -80 to +80 region; TATA+INR, genes having both TATA and INR at a fixed orientation and distance from each other (TATA box 15 to 30 nt upstream of the INR, 211 genes, 2.1%) as well as genes with a TATA and INR in any orientation and spacing within the -80 to +80 region (1358 promoters, 13.2%); INR only, genes with an INR element but no TATA box; None, genes with neither a TATA box nor an INR element in the -80 to +80 region. The number (No.) and percent (%) of genes of each category are given. (B) *S. cerevisiae* promoters (6165 genes) were searched for the presence of at least one TATA-8 element in the -150 to +1 region and/or one INR element in the -25 to +25 region and grouped into different promoter categories as in (A). The number (No.) and percent (%) of genes in each category are given.

A

Genes w/ Motif:	Binding Factor	Consensus	TATA only	TATA + INR	INR only	None	Total no. of genes
All	n.a.	n.a.	8.4% (865)	15.3% (1569)	30.3% (3107)	46.1% (4730)	10,271
M6	Sp1	GGGCGGR	6.6% (160)	4.2% (102)	26.3% (634)	62.8% (1514)	2410
M3	ELK-1	SCGGAAGY	4.1% (35)	5.7% (49)	29.6% (253)	60.5% (517)	854
M22	unknown	TGCGCANK	4.8% (34)	4.9% (35)	19.4% (138)	70.9% (505)	712

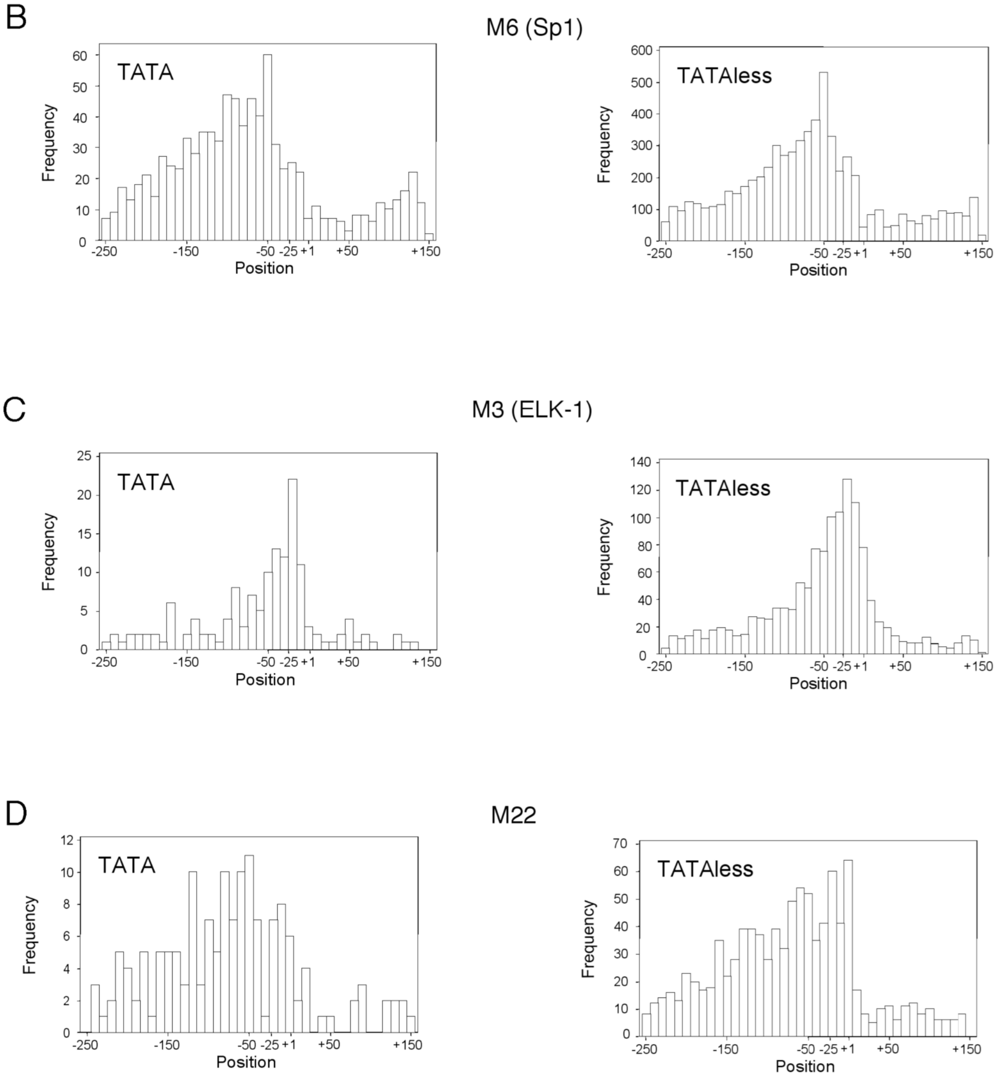

Fig 6.

Conserved promoter motifs selectively enriched in human TATA-less promoters. (A) The four categories of human promoters described in Fig. 5 (TATA only, TATA+INR, INR only, and None) were searched in the -250 to +150 region with the indicated consensus motifs from (Xie et al., 2005): M6 (an Sp1 binding motif), M3 (an ELK-1 binding motif), and M22 (binding factor unknown). The percent (%) and number (in parentheses) of genes with the indicated motifs that belong to each core promoter category are given and compared to all the genes (All). Frequency profiles of (B) M6 motifs, (C) M3 motifs, and (D) M22 motifs, within the -250 to +150 region of TATA-containing genes (2434 total) and TATA-less genes (7837 total) as defined in Fig. 5 are shown with a bin size of 10 nt.

A

Select Overrepresented GO Biological Processes

Biological Process	**% of genes (no.)**	**EASE score**
TATA only (594 genes annotated)		
Organogenesis	12.1% (72)	1.53E-05
Response to biotic stimulus	10.8% (64)	8.02E-05
TATA plus INR (module) (168 genes annotated)		
Nucleosome assembly	5.9% (10)	2.72E-07
Homophilic cell adhesion	7.1% (12)	6.51E-07
INR only (1941 genes annotated)		
Protein biosynthesis	7.34% (143)	2.05E-06
mRNA processing	2.4% (46)	2.27E-05
None (2287 genes annotated)		
Intracellular transport	8.05% (184)	5.51E-13
Cell growth and/or maintenance	35.7% (817)	1.80E-07
Protein metabolism	24.1% (552)	5.06E-05

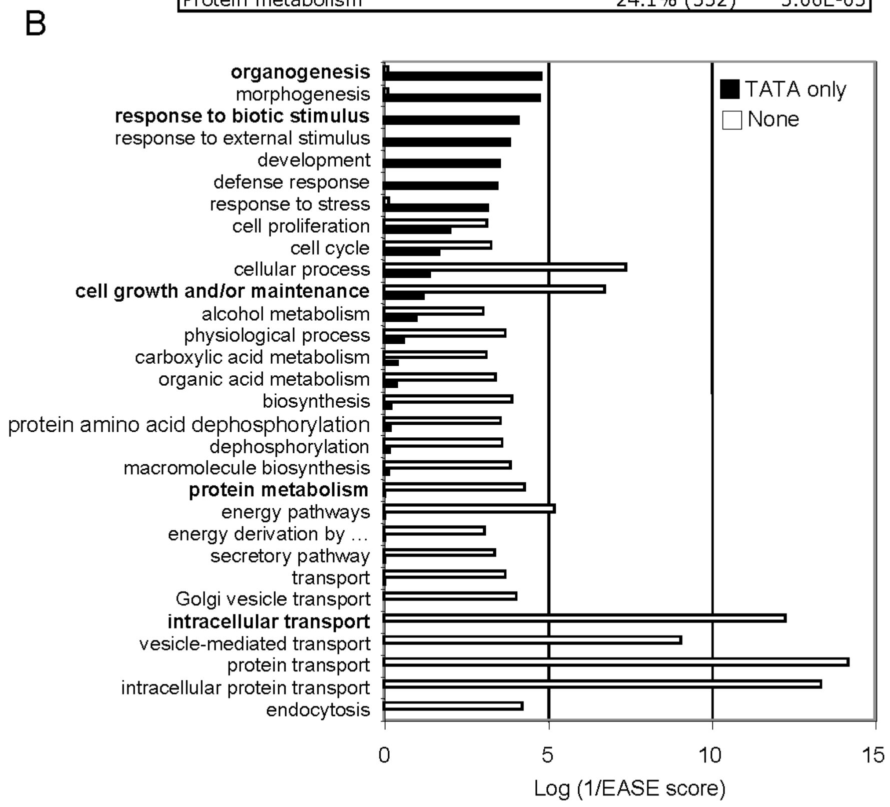

Fig 7.

Human genes in distinct core promoter categories are associated with different biological processes. (A) Shown are the most overrepresented Biological Processes from Gene Ontology (GO, http://www.geneontology.org/) for different core promoter categories: TATA only, INR only and None as defined in Fig. 5. TATA plus INR (module) is a subset of TATA+INR genes in which a TATA box is 15 to 30 nt upstream of an INR within the -80 to +80 window (see text for more details). Given is the percent (and number) of genes in a given promoter category that fall within a given Biological Process (total number of genes annotated for Biological Processes in a given promoter category is also indicated), with the corresponding EASE (Expression Analysis Systematics Explorer) score (http://apps1.niaid.nih.gov/david/; Hosack et al., 2003) which uses the upper bound of distribution of jackknife Fisher exact probabilities to distinguish enriched gene categories with respect to the entire DBTSS database. Shown are the largest non overlapping gene categories with EASE scores of 0.0001 or lower, that are specific to a given promoter category. (B) Histogram comparing the EASE scores of GO Biological Processes of the “TATA only” and “ None” categories. Shown are all categories for which the EASE score is <0.001 for either category. The Biological Process “energy derivation by oxidation of organic compounds” is truncated. Bold, Biological Processes shown in (A). The “INR only” category was not as specifically enriched compared to the “None” category, except for the Biological processes: Protein biosynthesis mRNA metabolism and oxidative phosphorylation (see supplementary Fig. S8). See supplementary Fig. S9 for a comparison of the Cellular Component of “TATA only” v. the “None” genes.