# Analysis of overrepresented motifs in human core promoters reveals dual regulatory roles of YY1

Hualin Xi, Yong Yu, Yutao Fu, et al.

*Genome Res.* 2007 17: 798-806
Access the most recent version at doi:10.1101/gr.5754707

Letter

# Analysis of overrepresented motifs in human core promoters reveals dual regulatory roles of YY1

Hualin Xi,[1] Yong Yu,[1] Yutao Fu,[1] Jonathan Foley,[2] Anason Halees,[1] and Zhiping Weng[1,2,3]

[1]*Bioinformatics Program, Boston University, Boston, Massachusetts 02215, USA;* [2]*Department of Biomedical Engineering, Boston University, Boston, Massachusetts 02215, USA*

A set of 723 high-quality human core promoter sequences were compiled and analyzed for overrepresented motifs. Beside the two well-characterized core promoter motifs (TATA and Inr), several known motifs (YY1, Sp1, NRF-1, NRF-2, CAAT, and CREB) and one potentially new motif (motif8) were found. Interestingly, YY1 and motif8 mostly reside immediately downstream from the TSS. In particular, the YY1 motif occurs primarily in genes with 5′-UTRs shorter than 40 base pairs (bp) and its locations coincide with the translation start site. We verified that the YY1 motif is bound by YY1 in vitro. We then performed detailed analysis on YY1 chromatin immunoprecipitation data with a whole-genome human promoter microarray (ChIP-chip) and revealed that the thus identified promoters in HeLa cells were highly enriched with the YY1 motif. Moreover, the motif overlapped with the translation start sites on the plus strand of a group of genes, many with short 5′-UTRs, and with the transcription start sites on the minus strand of another distinct group of genes; together, the two groups of genes accounted for the majority of the YY1-bound promoters in the ChIP-chip data. Furthermore, the first group of genes was highly enriched in the functional categories of ribosomal proteins and nuclear-encoded mitochondria proteins. We suggest that the YY1 motif plays a dual role in both transcription and translation initiation of these genes. We also discuss the evolutionary advantages of housing a transcriptional element inside the transcript in terms of the migration of these genes in the human genome.

[Supplemental material is available online at www.genome.org.]

The core promoter, consisting of ~100 bp flanking the transcription start site (TSS), plays an essential role in transcriptional initiation. It facilitates the assembly of the transcription initiation complex around the TSS. The TATA box is the best-characterized motif in this region (Smale and Kadonaga 2003). Its proper positioning is required to determine the starting point of the transcription; however, many promoters do not contain a TATA (Smale 1997), and the transcription initiation mechanisms for these promoters are not yet well understood. Several other core promoter motifs have been identified, with Initiator (Inr) being the best-studied example (Smale and Baltimore 1989; Javahery et al. 1994). It has the consensus YYAN(T/A)YY and is functionally similar to TATA in facilitating TFIID (TBP) binding. Beside Inr, a downstream promoter element called DPE was found in *Drosophila*. It occurs frequently in TATA-less promoters and appears to function cooperatively with Inr (Burke and Kadonaga 1996). Most recently, a new downstream core promoter element was identified by analyzing overrepresented motifs in *Drosophila* core promoter sequences and later verified experimentally (Ohler et al. 2002; Lim et al. 2004). Downstream core promoter motifs occur frequently in *Drosophila*; however, their occurrence in humans remains to be seen. In humans, Sp1 and CAAT box have been reported in several TATA-less promoters as the regulatory elements for transcription initiation (Huber et al. 1998; Mantovani 1998).

**[3]Corresponding author.**
**E-mail zhiping@bu.edu; fax (617) 353-6766.**
Article is online at http://www.genome.org/cgi/doi/10.1101/gr.5754707. Freely available through the *Genome Research* Open Access option.

Typical mammalian transcription regulatory motifs span only 6–12 bp and are degenerate. When the total base pair of input sequences is large, it is difficult to distinguish the real regulatory motifs in them from random short sequences. Fortunately, core promoters consist of only a short stretch of sequences flanking the transcription start sites. This drastically reduces the sequence search space. Recent efforts in large-scale sequencing of human full-length cDNAs have provided a unique opportunity to identify the precise positions of TSSs. In this study, we extracted high-quality human core promoter sequences (defined as −70 to +50 bp around the TSS throughout this study) from the Database of Transcription Start Sites (DBTSS, http://dbtss.hgc.jp) and analyzed them with the motif-finding program MEME (Bailey and Elkan 1994; Bailey et al. 1997). TATA was found in only 26% of these promoters. In addition to TATA and Inr, we identified several overrepresented motifs. Interestingly, some of these motifs, although generally overrepresented in the entire set of core promoters, are underrepresented in TATA-containing promoters. With all of the overrepresented motifs identified in this study, >50% of the TATA-less core promoters can be accounted for, indicating an improved understanding of the general transcriptional initiation mechanism.

One identified motif was particularly interesting. It matched the YY1-binding consensus; however, it occurred downstream from the TSS and often overlapped with the translation start site. We demonstrated experimentally that it could be recognized by YY1 in vitro. The similarity between this YY1 motif and the Kozak sequence (Kozak 1984; Kozak 1987b) led us to question whether the YY1 motif was merely a special Kozak sequence and its ability to bind YY1 was coincidental. Two observations we


17:798–806 ©2007 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/07; www.genome.org

made suggest otherwise: The YY1 motif was evolutionarily more conserved than Kozak and most genes with the YY1 motif had extremely short 5′-UTRs (<40 bp). In addition, the raw experimental data on YY1 chromatin Immunoprecipitation with a whole-genome human promoter microarray (ChIP-chip) was made available to us (B. Ren, pers. comm.). We performed a detailed analysis on the ChIP-chip data and uncovered a series of striking parallels between the aforementioned results on the YY1 motif and those on the YY1 ChIP-chip data. Both consistently suggested two distinct binding modes of YY1: one on the plus strand downstream of the TSS and the other on the minus strand upstream of the TSS, with the former often overlapping the translation start site and the latter transcription start site. We supply multiple lines of computational evidence to argue distinct regulatory mechanisms by the two binding modes and discuss the evolutionary implication of the dominant downstream-of-TSS mode. As little is known about downstream core promoter elements in humans, our finding here could potentially represent a new mechanism for transcription regulation in eukaryotes.

## Results

We first analyzed overrepresented motifs found in core promoters in terms of their consensus, positional specificity, co-occurrence with TATA, and evolutionary conservation. Then, one overrepresented motif, YY1, was characterized in great detail by using both experimental and computational approaches.

### Overrepresented core promoter motifs

Based on the quality assessment of TSS mapping in DBTSS (see Supplemental Fig. 1), 723 core promoter sequences with >15 mapped cDNA sequences were selected for MEME motif analysis. Inr, known to be located at the +1 position, was between –70 and +30 in 90% of these 723 sequences, indicating accurate mapping of the TSSs. The top 15 scoring motifs found by MEME were recorded. Among them, five highly degenerated motifs were omitted from further analysis (Supplemental Fig. 3). The remaining 10 motifs are listed in Table 1. The position-specific scoring matrices (PSSMs) of these motifs were compared with all matrices in the TRANSFAC database by using the MALIGN algorithm (Haverty et al. 2004a). Nine motifs matched TRANSFAC matrices, while motif8 appeared to be novel (the first column of Table 1). The known motifs included core promoter motifs (TATA, Inr) and several other well-studied motifs (CAAT box, Sp1, NRF-1, NRF-2 [also known as GABP], and CREB). The second-best scoring motif matched the previously reported YY1-binding profile (Shrivastava and Calame 1994) in the reverse direction.

All 10 motifs showed positional specificity with respect to the TSS (Table 1). TATA and Inr, as expected, showed extremely strong positional specificity. Similar levels of positional specificity were also observed for YY1 and motif8 and both peaked immediately downstream from the TSS. Co-occurrence of these motifs with TATA was analyzed in the 723 promoter sequences (Supplemental Table 1). YY1, NRF-1, NRF-2, and motif8 showed significantly lower-than-expected co-occurrence with TATA. In the extreme case of NRF-2, only two cases of co-occurrence were observed as opposed to the expected 13 cases. Thus, these motifs might assist transcription initiation in TATA-less genes. The statistics for other motifs including CAAT and CREB were insignificant, possibly due to their lower abundance in the sequence set. All 10 motifs found in humans were also overrepresented in a set of 1849 high-quality mouse core promoter sequences from DBTSS with >10 mapped cDNA sequences. As conservation often suggests functional importance, overrepresentation observed in both humans and the mouse supports the functional relevance of these motifs in core promoters.

### The YY1 motif overlaps with the Kozak sequence

Of the 723 high-quality DBTSS promoters, 54 contained the YY1 motif. For 44 of them, the YY1 motif overlapped with the translation start site, with the ATG nucleotides in the motif coinciding with the initiation codon. The sequence surrounding a translation start site, commonly referred to as the Kozak sequence (Kozak 1984; 1987b), facilitates the assembly of the ribosome around the mRNA molecule during translation initiation. De-

**Table 1.** Summary of overrepresented motifs identified from the MEME analysis

| Motif | Logo | Bits | # in 723 high quality promoters | # in 10577 DBTSS promoters | Positional Distribution relative to TSS[a] | P-value |
|---|---|---|---|---|---|---|
| 1 (TATA) | | 13 | 232 | 2634 | | 6.9E-186 |
| 2 (YY1) | | 17 | 54 | 677 | | 4.1E-55 |
| 3 (Inr) | | 17 | 47 | 248 | | 8.0E-46 |
| 4 (Sp1) | | 13 | 112 | 1785 | | 1.8E-33 |
| 5 (Sp1 reverse) | | 14 | 57 | 1204 | | 7.3E-18 |
| 6 (NRF1) | | 15 | 51 | 561 | | 2.1E-15 |
| 7 (NRF2) | | 16 | 39 | 562 | | 6.9E-19 |
| 8 (motif8) | | 15 | 49 | 764 | | 3.6E-11 |
| 9 (CAAT) | | 19 | 16 | 142 | | 1.2E-08 |
| 10 (CREB) | TGACGTCA | 17 | 13 | 408 | | 4.6E-03 |

Known motifs are indicated in parentheses. If a motif occurs multiple times in the same promoter, the position of the best match was used in calculating the positional distribution.
[a]Number of promoters containing the motif in the 10,577 unique DBTSS human promoters.

spite the positional overlap of the YY1 motif and the Kozak sequence, three differences were observed between them. First, the optimal Kozak sequence 5′-CRCC**AUG**G-3′ (Smith et al. 2005) is different from the YY1 consensus in several positions (CAAGATGGCGGC, differences underlined). In particular, the four positions at the 3′ end of the YY1 consensus (CGGC) are not required for the Kozak sequence. This strongly suggests that the YY1 motif may have other regulatory functions. Second, the human Kozak sequence is much more degenerate than the YY1 motif. In a recent study, when the nucleotides flanking the AUG start codon for 22,208 human genes were aligned, weak consensus was observed only for the –3 and +4 positions (5′-NRNN**AUG**G-3′) (Smith et al. 2005). In contrast, the YY1 motif has high information content at most positions (Table 1; the YY1 PSSM is provided in Supplemental Materials). Third, YY1 sites are more evolutionally conserved between humans and the mouse than the Kozak sequence (see below).

### The occurrence of the YY1 motif correlates with short 5′-UTR

Most of the aforementioned 44 genes with the YY1 motif at their translation start sites have short 5′-UTRs. The distribution of 5′-UTR lengths among the 723 genes that correspond to the high-quality DBTSS promoter set is shown in Figure 1 (empty boxes). Interestingly, this is a bimodal distribution with two peaks centered on 20 and 60 bp, respectively. A similar distribution is observed for genes with TATA (gray boxes). However, 85% of the genes with the YY1 motif have 5′-UTRs shorter than 40 bp (black boxes), coinciding largely with the first peak at 20 bp in the distribution of all genes. The distribution of the YY1-containing genes differs significantly from that of TATA genes, with a *P*-value of $3.0 \times 10^{-18}$ according to the Wilcoxon rank sum test.

### YY1 sites are conserved between human and mouse

Searching against DBTSS, we found 168 genes with a 5′-UTR shorter than 30 bp and a YY1 motif spanning from −4 to +8 bp relative to the translation start site. These 12 nucleotide positions are significantly more conserved between humans and the mouse in the 168 genes than in a set of randomly chosen genes that do not contain YY1 sites (Fig. 2) ($P = 0.001$ according to the paired Student's *t*-test). In particular, 15%–20% fewer mismatches were observed at the −3, −2, and −1 positions for the former set, while much smaller differences were observed for positions outside of the YY1 site (−12 to −5 and +9 to +15; data not shown). The additional conservation observed in the YY1 motif-matching region around the translation start site further suggests its function beyond a translation initiation motif.

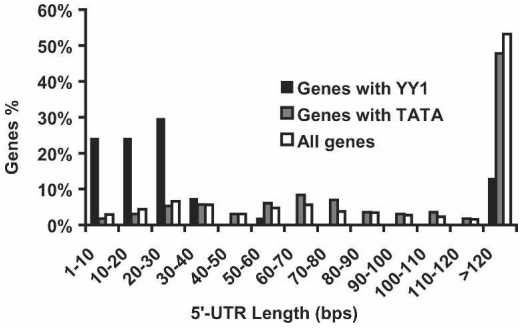


**Figure 1.** The distribution of the 5′-UTR length in genes with a YY1 motif match (black), TATA (gray), and all genes in the 723 high-quality DBTSS promoters set.

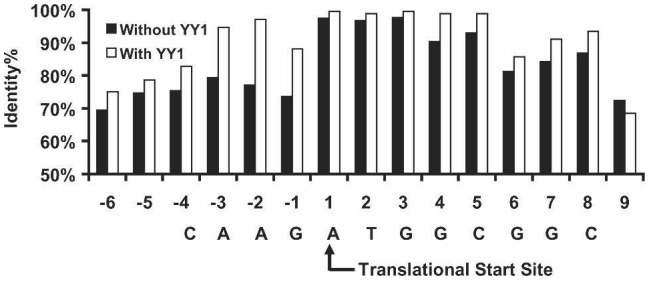


**Figure 2.** Sequence conservation between human and mouse for the positions around the translation start sites with (white) or without (black) a YY1 motif match.

### Predicted YY1 sites are recognized by the YY1 protein in vitro

The consensus of the YY1 motif (CAAGATGGCGGC) derived from MEME analysis matches perfectly the reverse complement of the previously reported YY1-binding site consensus GC CATCTTG (Shrivastava and Calame 1994). To verify experimentally whether our predicted YY1 sites were indeed capable of YY1 binding, eight 25-bp-long promoter sequences, each centered on a predicted YY1 site, were randomly selected for electrophoretic mobility shift assay (EMSA). The results of three sequences are shown in Figure 3; the rest of the results and the experimental protocol are found in Supplemental Materials. SC-2533, a known YY1-binding sequence (Hariharan et al. 1991), was used as the positive control. The negative control sequence was derived from one of the eight sequences by introducing four mutations at the putative YY1-binding site. For further comparison with a known YY1-binding site in L1 elements, a sequence that covers the site in the L1 5′-UTR was also included in the experiment. For all eight sequences, a shifted band was observed when the radiolabeled oligomers were incubated with HeLa cell extract. A supershift band was observed after an anti-YY1 antibody was added, further confirming that the shift was specifically caused by YY1 binding. No radioactive band was observed for the negative control, indicating that the mutations completely abolished YY1 binding. The gel mobility-shift patterns of all eight sequences were identical to that of the L1 sequence.

### Genome-wide chromatin immunoprecipitation with HeLa cells indicates that YY1 binds to YY1 motifs in two distinct modes

To study YY1-binding sites in living cells, we analyzed YY1 ChIP-chip data on a human genome-wide promoter array (B. Ren, pers. comm.). A total of 24,135 promoters were probed by the array. Among them, 765 showed significant YY1 binding (called ChIP hits), indicated by ChIP signal two standard deviations above the mean. We searched the YY1 PSSM (generated with MEME on the 723 high-quality DBTSS promoters as described above) (Table 1) against all 24,135 promoters. Only the regions covered by the probes were searched, typically −1300 to +200 bp around the TSS. A 10-fold enrichment of the YY1 motif was found on the plus strand of the ChIP hits when compared with non-hits (Table 2A). Such a highly significant enrichment ($P = 2.2 \times 10^{-16}$) cannot be trivially explained by a general enrichment of genuine core promoters in the ChIP hits, as no or only slight enrichments (one- to twofold) were observed for three other core promoter motifs, TATA, NRF-2, and CREB (Table 2C,D,E). Surprisingly, a lower but still highly significant fivefold enrichment of the YY1 motif was also observed on the minus strand of the ChIP hits

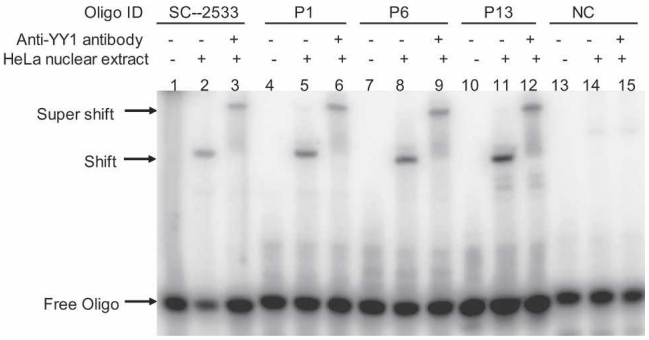


**Figure 3.** EMSA of YY1 binding to the predicted YY1 motif. Oligo SC-2533 was used as the positive control and NC stands for negative control. HeLa nuclear extract was the source for the YY1 protein. Anti-YY1 mouse IgG was used for the super-shift reactions. The mobility shift and super shift of the oligos demonstrate the specific binding of YY1 to YY1 motif. The EMSA shown was a phosphorimage of the gel.

(Table 2B). Of the 723 high-quality DBTSS promoters, 535 were probed by ChIP-chip. These included 47 promoters that contained the YY1 motif, of which 42% were ChIP hits, a fivefold enrichment over the promoters that do not contain the YY1 motif (Table 2F) ($P = 2.9 \times 10^{-12}$).

When analyzed for the positional distribution of the YY1 motif in these 765 ChIP hits, different positional specificities were observed on plus and minus strands. On the plus strand, the YY1 motif peaks at +1 to +20 bp downstream the TSS, while on the minus strand it peaks at −40 to +1 bp, mostly upstream from the TSS (Fig. 4; Supplemental Fig. 5). Furthermore, when we averaged the ChIP signal of all ChIP hits that contain YY1 sites on the plus strand, we observed a single peak at +30 to +80 bp downstream of the TSS (thick black curve in Fig. 5), while the average ChIP intensity generated with ChIP hits that contain the YY1 site on the minus strand had a broader peak at −60 to −10 bp upstream the TSS (thin black curve in Fig. 5). This clear correspondence between the positional distributions of the YY1 motif and the YY1 ChIP signal indicates that YY1 is capable of binding to the predicted motif in both orientations in living cells, and these are two distinct binding modes.

Furthermore, the genes corresponding to the ChIP hits with YY1 sites on the plus strand tend to have short 5′-UTRs: 34% (102) of these 294 genes have 5′-UTRs shorter than 40 bp (Fig. 6). This trend is even more pronounced for ChIP hits with YY1 sites overlapping the translational start site on the plus strand: 78% (87) of these 109 genes have 5′-UTRs shorter than 40 bp. In contrast, only 10% of the genes corresponding to ChIP hits with YY1 sites on the minus strand have short 5′-UTRs, at the same level as random. The plus-strand mode is consistent with, although with a lower percentage of short 5-UTRs, our earlier observation in the 723 high-quality DBTSS promoters, where 85% of the genes with downstream YY1 sites have 5′-UTRs shorter than 40 bp (Fig. 1). The higher percentage for the DBTSS promoters likely reflects their higher TSS mapping accuracy than the promoters on the microarray, for which the TSS coordinates were obtained from RefSeq. Indeed, when the TSS coordinates from DBTSS were used (available for 188 of the 294 ChIP hits with YY1 sites on the plus strand), the percentage of genes with <40 bp 5′-UTRs rose to 47% (Supplemental Fig. 6). The positional specificity of the YY1 motif also became sharper when the TSS coordinates from DBTSS were used (cf. Supplemental Figs. 5 and 4).

## YY1 PSSM can predict accurately YY1-binding promoters identified by ChIP

Four hundred (52%) of the 765 ChIP hits contained a YY1 site on the plus or minus strand. For the remaining 365 ChIP hits, no

**Table 2.** Enrichment of motifs in ChIP hits

**A. YY1 motif on +strand**

| | – YY1 | + YY1 | Occurrence |
|---|---|---|---|
| Non-hit | 22106 | 806 | 3.5% |
| ChIP hit | 471 | 294 | 38.4% |
| Hit rate | 2.1% | 26.7% | $P(\chi^2) < 2.2 \times 10^{-16}$ |

**B. YY1 motif on −strand**

| | – YY1 | + YY1 | Occurrence |
|---|---|---|---|
| Non-hit | 22248 | 670 | 2.9% |
| ChIP hit | 637 | 128 | 16.7% |
| Hit rate | 2.8% | 16% | $P(\chi^2) < 2.2 \times 10^{-16}$ |

**C. TATA (TATAAA) motif**

| | – TATA | + TATA | Occurrence |
|---|---|---|---|
| Non-hit | 16458 | 6460 | 28.2% |
| ChIP hit | 569 | 196 | 25.6% |
| Hit rate | 3.3% | 2.9% | $P(\chi^2) = 0.13$ |

**D. CREB (TGACGTCA) motif**

| | – CREB | + CREB | Occurrence |
|---|---|---|---|
| Non-hit | 22261 | 657 | 2.9% |
| ChIP hit | 719 | 46 | 6% |
| Hit rate | 3.1% | 6.5% | $P(\chi^2) = 3.8 \times 10^{-7}$ |

**E. NRF-2 (GGAAGTG) motif**

| | – NRF-2 | + NRF-2 | Occurrence |
|---|---|---|---|
| Non-hit | 19511 | 3407 | 14.9% |
| ChIP hit | 614 | 151 | 19.7% |
| Hit rate | 3.1% | 4.2% | $P(\chi^2) = 1.8 \times 10^{-4}$ |

**F. Hit rate enrichment in the 723 high-quality DBTSS promoters**

| | – YY1 | + YY1 | Occurrence |
|---|---|---|---|
| Non-hit | 449 | 27 | 5.7% |
| ChIP hit | 39 | 20 | 33.9% |
| Hit rate | 7.9% | 42% | $P(\chi^2) = 2.9 \times 10^{-12}$ |

Contingency tables were used to determine whether the YY1 motif was enriched on either plus or minus strands of probe-covered regions that correspond to ChIP hits. The probe-covered regions without significant YY1-binding signal (abbreviated as "non-hit" in the tables) were used as background. *P*-value cutoff of 0.00001 was used to determine YY1 motif match. $\chi^2$ statistics was used to test whether the enrichments were significant. Several other overrepresented core promoter motifs including TATA box, CREB, and NRF-2 were used as controls. For these motifs, exact string match to their consensus (shown in parentheses) was used to determine motif matches. Enrichment of YY1 sites is highly significant on both plus and minus strands, while only slight enrichment is observed for CREB and NRF-2, and no enrichment is observed for TATA box. *F* shows significant enrichment of hit rate in high-quality DBTSS promoter with the YY1 motif.

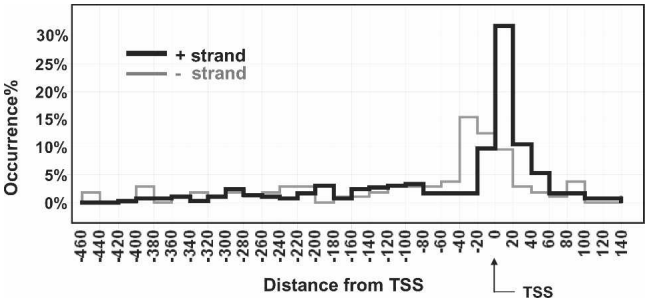


**Figure 4.** Positional distribution of YY1 sites relative to the TSS for ChIP hits. Thick black line: distribution on the plus strand. Thin gray line: distribution on the minus strand. Different positional specificities were observed for YY1 sites on plus and minus strands.

other obvious motif was found to be enriched in the probe-covered regions when a MEME search was carried out. Some of these promoters might contain weak YY1 sites, as we used a stringent *P*-value cutoff ($<10^{-5}$) for matching the YY1 PSSM. Indeed, when we relaxed the cutoff slightly to $4\times10^{-5}$, 534 (70%) of the 765 ChIP hits contained YY1 sites on either strand. The average ChIP signal for these 365 promoters seems to peak even further downstream from the TSS (black curve in Fig. 6). This suggests that additional proteins might be involved in assisting the binding of YY1.

Also of interest are the false-positive rates of the YY1 PSSM at various sensitivity levels. We scanned the 10,000 promoters with the weakest YY1 ChIP signals for YY1 sites at a series of *P*-value cutoffs. The percentage of promoters that contained at least one YY1 site was defined as the false-positive rate and plotted against the sensitivity at the same *P*-value cutoff (Fig. 7). On average, there are twice as many YY1 sites on the plus strand (the curve labeled with "YY1") than on the minus strand (the curve labeled with "YY1r," where r indicates reverse). When both strands are considered (the curve labeled with "YY1|YY1r"), a genome-wide specificity of 96.5% is achieved at a sensitivity of 52%, indicating that the YY1 PSSM is highly predictive of YY1 binding in living cells. One might expect that top ChIP hits can be used to optimize the YY1 PSSM in an iterative fashion. We attempted this with the ROVER algorithm (Haverty et al. 2004b); however, it converged after one round of iteration. The resulting PSSM did not lead to improved performance (the curve labeled "YY1o," where o stands for optimization; the training promoters were excluded during the sensitivity calculation). This indicates that the YY1 PSSM generated with the DBTSS promoters were already optimal for predicting YY1 binding in living cells.

### Biological functions of genes with YY1 sites

We performed functional enrichment analysis on the genes corresponding to the 765 ChIP hits after dividing them into three sets: genes with YY1 sites on the plus strand (220 annotated), genes with YY1 sites on the minus strand (87 annotated), and genes without YY1 sites (189 annotated). For set 3, a loose *P*-value cutoff (<0.0001) was used to exclude genes with possible YY1 motif matches on either strand. We searched for the enrichment of any GO terms in each set of genes using GoStat (Beissbarth and Speed 2004). The 10,000 genes with the weakest ChIP signal were used as the background.

Genes with YY1 sites on the plus strand are highly enriched in the GO term mitochondrion (Table 3). Specifically, 36 genes in set 1 are encoded in the nuclear genome, but their protein products function in the mitochondria, including mitochondrial membrane proteins, enzymes, and a large number of ribosomal proteins. Gene sets 2 and 3 are only marginally enriched in this term. For the genes with YY1 sites on the minus strand, several ribosome-related terms were highly enriched; however, most of them were also among the most enriched terms in gene sets 1 and 3. Only the genes without YY1 sites were enriched in the term RNA binding. Genes annotated with this term are different from the ribosomal proteins found in gene sets 1 and 2; the former are involved in RNA post-transcriptional processing or RNA metabolism, including polyadenylation factors, splicing factors, and RNAases. We note that these enriched functional terms collectively account for only ~20% of the annotated YY1 target genes. Overall, the genes whose promoters are bound by YY1 in HeLa cells seem to be involved in a diverse set of cellular functions and processes.

## Discussion

### Accurate TSS mapping and focused search in core promoters led to the discovery of high-quality transcription-factor binding motifs

Transcription-factor binding sites are in general short and degenerate. The difficulty in mapping the TSS accurately further complicates the effort in finding core promoter motifs. Our strategy was to use only sequences with accurate TSS locations. This allowed us to interrogate the short stretch of sequences (from −70 to +50 bp) that were most likely to cover the real core promoters. Our approach boosted the core promoter element signal and decreased background noise. We identified eight previously known motifs (TATA, Inr, YY1, Sp1, CAAT, NRF-1, NRF-2, and CREB) and one potentially novel motif. FitzGerald et al. analyzed the positional distribution of all 8-mers around the putative TSSs of 13,010 human genes and reported eight overrepresented motifs and the Kozak sequence (FitzGerald et al. 2004). We also found six of these eight motifs, missed USF and Clus1, but gained YY1, Inr, and the potentially novel motif8. The two studies used completely different approaches. Most importantly, we provided several lines of evidence indicating that YY1 is a transcription motif, although it often overlaps with the Kozak sequence of genes with short 5′-UTRs.

We estimated that >90% of the 723 high-quality human promoters studied would have their actual TSS located within the

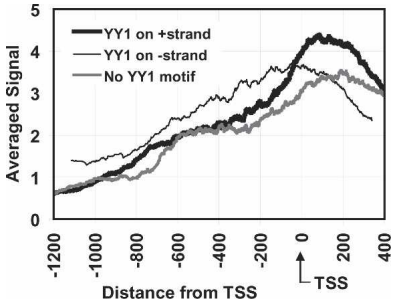


**Figure 5.** Positional distribution of averaged ChIP signal relative to the TSS for ChIP hits. Thick black curve: ChIP hits with a YY1 motif match on the plus strand (*P* <0.00001). Thin black curve: ChIP hits with a YY1 motif match on the minus strand (*P* <0.00001). Thick gray curve: ChIP hits without a YY1 motif match on either strand.

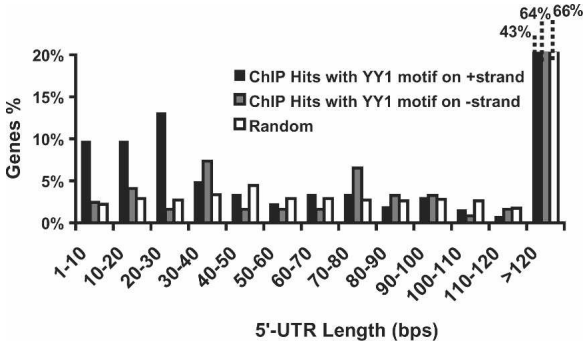


**Figure 6.** 5′-UTR lengths of genes that correspond to ChIP hits. 5′-UTR length distributions of ChIP hits with a YY1 motif hit on the plus strand (black bars) and on the minus strand (gray bars) were compared with the distribution of randomly selected genes (white bars) probed in the YY1 ChIP-chip experiment.

120-bp window. The recent CAGE study by the RIKEN group indicates that CpG promoters tend to have a broad TSS distribution, with variation in TSS positions up to 50 bp (Carninci et al. 2006). Thus, the promoters that do not have an Inr at the +1 position could correspond to alternative TSSs, or could be due to inaccurate TSS mapping. This is unlikely to affect the quality of the PSSMs derived in this study as they are highly enriched, but may slightly affect the positional specificity analysis on these motifs.

### The predicted YY1 motif was confirmed by EMSA and ChIP-chip data

Our results show that the predicted YY1 motif is bound by YY1 in vitro and the PSSM is highly predictive of YY1 binding in living cells. It is interesting that the PSSM derived with MEME could not be further improved by using top ChIP hits. In contrast, we used a similar approach in an earlier study to refine the ab initio PSSM of the p53 motif and drastically improve its predictive accuracy in the human genome (Wei et al. 2006).

We were prompted to investigate the distribution of YY1 sites in nonpromoter regions, as the probes on the promoter array covered only 36 M base pairs in total. The YY1 PSSM could detect 50% of ChIP hits at the 97% specificity of rejecting non-YY1 promoters genome-wide. At the same stringency cutoff, 89,000 YY1 sites were found in the 1.5 G-base-pair nonrepetitive portion of the human genome. These many sites are expected at the corresponding 97% specificity, indicating that there is no under-representation of YY1 sites outside promoters. Nonetheless, there is a strong enrichment of YY1 sites in the ±100-bp window around the TSS: 6.3 folds on the plus strand and 3.4 folds on the minus strand. This is compared with the even greater enrichment of YY1 sites observed for ChIP hits vs. non-hits: 11 folds on the plus strand and 5.8 folds on the minus strand (Table 2A,B). Thus, the YY1 motif can explain YY1 binding in living cells, although not entirely, and there are other contributing factors, possibly chromatin structure or the interaction of YY1 with other transcription regulators.

### ChIP-chip data reveals two binding modes of YY1, both consistent with YY1 motif enrichment

YY1 has been studied extensively since it was identified in 1991 (Shi et al. 1991). It is a multifunctional protein that can repress, activate, or initiate transcription, depending upon the context of its binding site and the presence of other regulatory proteins (Shi et al. 1997). Most of the previously reported YY1-binding sites are upstream of the TSS, and genes with downstream sites primarily include ribosomal protein genes and L1 elements (Riggs et al. 1993; Safrany and Perry 1995; Cole and Gaston 1997; Li et al. 1997; Shi et al. 1997; Thomas and Seto 1999; Athanikar et al. 2004). YY1 can either be activating or repressive, irrespective of whether it binds upstream or downstream of the TSS. YY1 was first reported to bind the sequence surrounding the TSS of the AAV P5 promoter, and this interaction was shown to initiate transcription in vitro (Shi et al. 1991). Subsequently, a handful of promoters have been reported to be in this category (Shi et al. 1997). Consistent with this function, YY1 has been reported to interact with several components of the basal transcription apparatus, including TBP, GTF2B (formerly TFIIB) and TAF7 (formerly TAFII55) (Thomas and Seto 1999). More recently, it is suspected to interact with the polycomb group proteins (e.g., Suz12) and participate in gene silencing (Srinivasan and Atchison 2004).

This study represents a systematic assessment of YY1-binding promoters genome-wide in HeLa cells. Our results indicate that the promoters of >3% of the human genes in HeLa cells have detectable YY1 binding with ChIP-chip. We do not observe significant overlap between the ChIP hits of YY1 and SUZ12 in ENCODE regions (The ENCODE Project Consortium 2007). Instead, there is a significant overlap between the ChIP hits of YY1 and TAF1 (a component of TBP) in HeLa cells, with the fold of enrichment even higher than that between TAF1 and POLR2A (16 vs. 13 folds; ENCODE data). These results suggest that YY1 binding frequently indicates active transcription. We extrapolate to speculate that YY1 is more often an activator than a repressor.

At least 50% and likely 70% of ChIP hits can be accounted for by the YY1 motif. These ChIP hits are divided into two classes: 2/3 of them contain YY1 sites on the plus strand and 1/3 of them contain YY1 sites on the minus strand. For these two classes, a clear difference in positional distribution of YY1 sites was observed (Fig. 4). Similar difference in positional specificity was also observed for the two classes in average ChIP signal (Fig. 5). The dominance of downstream plus-strand binding mode and the overlap of the YY1 site and the translational start site are novel

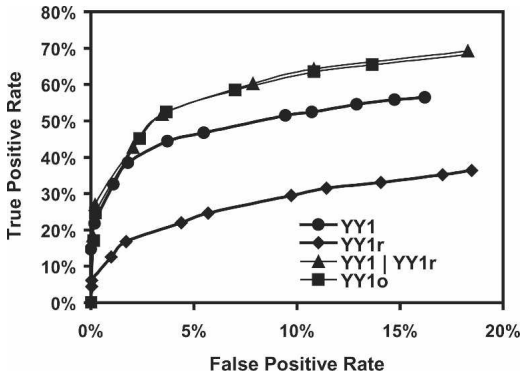


**Figure 7.** The YY1 PSSM accurately predicts ChIP hits. The YY1 PSSM was searched against probe-covered regions of the 765 ChIP hits (true positives) and the 10,000 probe-covered regions with the lowest ChIP signal (true negatives). True positive rates were plotted against the false positive rates at various *P*-value cutoffs. (YY1) YY1 motif on the plus strand; (YY1r) YY1 motif on the minus strand; (YY1/YY1r) YY1 motif on either the plus or the minus strand; (YY1o) YY1 motif after optimization in ROVER algorithm on either the plus or the minus strand.

**Table 3. Enrichment of GO terms in ChIP hits with YY1 sites on the plus strand, the minus strand, or neither strands**

| | # of annotated genes | Enriched GO term | GO term id | # of genes with GO term | *P*-value |
|---|---|---|---|---|---|
| ChIP hits with YY1 motif on +strand | 220 | Mitochondria | GO:0005739 | 36 | $1.23\times10^{-41}$ |
| | | Structural constituent of ribosome | GO:0005840 | 18 | $1.88\times10^{-15}$ |
| ChIP hits with YY1 motif on −strand | 87 | Structural constituent of ribosome | GO:0005840 | 11 | $1.87\times10^{-10}$ |
| ChIP hits without YY1 motif | 189 | RNA binding | GO:0003723 | 34 | $1.17\times10^{-21}$ |

The 10,000 genes with the lowest ChIP signal were used as the background. *P*-values were determined by the GoStat program (Beissbarth and Speed 2004).

findings in this work. When occurring on the minus strand, YY1 sites often overlap with the TSS, at the same location as Inr. This is consistent with the previously reported initiator function of YY1. Our results further previous studies by indicating that YY1 may initiate the transcription of many more human genes than previously known. The list of 18 genes for which the YY1 site overlaps Inr is provided in Supplemental Materials.

When occurring on the plus strand, a significant portion of the YY1 sites overlap with the Kozak sequence, in which case the associated genes frequently have short 5′-UTRs (<40 bp). In contrast, the ChIP hits with YY1 sites on the minus strand are not enriched with genes with short 5′-UTRs (Fig. 6). Thus, we propose that these downstream plus-strand YY1 sites correspond to a distinct regulatory mechanism, different from the upstream minus-strand sites, or the downstream plus-strand sites in genes with longer 5′-UTRs. The former may also impact translation efficiency. The dual regulatory roles of these sites on both transcription and translation may account for their greater extent of evolutionary conservation than Kozak sequences (Fig. 2). In the next section we speculate on the biological function and evolutionary advantages of these sites.

### Biological functions of YY1-regulated genes and speculation on the evolutionary role of YY1

YY1 is known to regulate the expression of many human genes (Li et al. 1997; Shi et al. 1997). In this study, we report 765 genes whose promoters are bound by YY1 in HeLa cells; these include a large number of genes that were not previously known to be regulated by YY1. They seem to be involved in a variety of cellular functions and processes with strong enrichment of nuclear-encoded mitochondrial genes, ribosomal protein genes, and genes involved in RNA processing.

Most strikingly, many nuclear-encoded mitochondria genes have downstream YY1 sites and they tend to have short 5′-UTRs. A similar arrangement was also observed in L1 elements (Becker et al. 1993), the most abundant repetitive element in the human genome, covering roughly 15% of the genome (Smit 1996). The YY1-binding consensus in L1 elements is identical to that found in this study. It also locates immediately downstream the TSS with a peak location between +10 and +20, similar to that of YY1 motif (Lavie et al. 2004). One study showed that YY1 is required for accurate transcription initiation and full-length 5′-UTR (Athanikar et al. 2004).

There are several intriguing questions associated with the YY1 motif when occurring downstream the TSS. What is the advantage of a downstream core promoter motif compared with a typical upstream one? What is the advantage of having the translation start site overlapping with YY1 motif? Is there any biological significance between the striking similarity of downstream YY1 sites in genes and L1 elements?

It is well known that short 5′-UTRs are indicative of efficient translation initiation (Kozak 1987a), as RNA secondary structures and upstream AUG codons are less likely to occur. Our analysis on downstream YY1 sites indeed revealed an enrichment of highly expressed ubiquitous genes, for example, ribosomal protein genes and nuclear-encoded mitochondria genes. We speculate that the binding of YY1 can help accurate positioning of transcription initiation and thus ensure the intact 5′-UTRs for these genes, just like its function on L1 elements (Athanikar et al. 2004). Moreover, these YY1 sites may be especially good Kozak sequences and further improve translation efficiency.

There might be a resemblance between the evolutionary histories of L1 elements and nuclear-encoded mitochondria genes. L1 elements move about the genome via a RNA intermediate by a process termed retrotransposition. Having a downstream core promoter element buried inside the 5′-UTR gives an L1 element the advantage of carrying along its own core promoter during transposition and maintaining an intact core promoter after inserting into a new site in the genome. Nuclear-encoded mitochondrial genes were originally encoded by the mitochondria genome and, at some point of evolution, transferred into the nuclear genome. Little is known about the precise mechanism of the mitochondrial gene transfer, except that this is a highly inefficient process and either RNA or DNA could serve as the transfer intermediate (Blanchard and Lynch 2000). We speculate that having a downstream YY1 site would lend substantial evolutionary advantage to the transfer via an RNA intermediate, because YY1 would ensure the full-length 5′-end of this RNA and thus preserve the transcription regulatory region, which would exert substantial survival advantage upon arrival in the nuclear genome.

## Methods

### Data sources

Various human promoter data sets used in this study are defined and compared in Supplemental Figure 2. The coordinates of human TSSs were downloaded from DBTSS (http://dbtss.hgc.jp; version 4.2). Core promoter sequences from –70 to +50 from the TSS were extracted from UCSC genome browser (http://genome.ucsc.edu) by using genome assembly hg16 and mm3. The accession numbers of the 723 high-quality human promoters are provided in the Supplemental Materials.

### Identification of overrepresented motifs

Overrepresented motifs were identified by using MEME (Bailey and Elkan 1994; Bailey et al. 1997) with the "-zoop" option, which indicates "zero or one occurrence per sequence," and motif width set to be between 6 and 15 bp. The 723 high-quality DBTSS promoters were used for this analysis. The top 15 motifs were obtained. For each of these, the positional specific scoring

matrix (PSSM) generated by MEME was searched against the TRANSFAC database using the MALIGN algorithm (Haverty et al. 2004a).

### Positional distribution of overrepresented motifs

Each motif was searched using MAST (Bailey and Elkan 1994; Bailey et al. 1997) against the –500 to +200 region of all 10,577 human promoters in DBTSS. A *P*-value cutoff between $10^{-5}$ and $10^{-6}$ was used for each motif to determine the matches (with the exception of TATA, where a 0.01 cutoff was used due to much shorter consensus sequence). Then, the distances between the matches and the TSS were calculated and the number of matches in every 25-bp interval from the TSS was plotted.

### Conservation of YY1 motif between human and mouse

A total of 508 human genes in DBTSS with 5′-UTR length shorter than 30 bp were identified and separated into two sets, 168 with YY1 sites and 340 without YY1 sites at the translation start site. The sequences covering −12 to +12 bp around the start codon were extracted. The corresponding sequences in the mouse genome were obtained from the UCSC genome browser. The conservation score at each position was calculated as the total number of matches divided by the total number of sequences.

### Defining hits in YY1 ChIP-chip experiment

The raw experimental data from the YY1 ChIP-chip experiment was made available to us (B. Ren, pers. comm.). A detailed description of the ChIP-chip experiment design described by Ren and colleagues (L. Shen, K. Wang, S. Agarwal, B. Ren, and W. Wang, in prep.). Briefly, ~1500 bp (−1200 to +300), each covering of a set of 24,135 promoters, were extracted from human genome (build 35, HGS17). This minimal set of promoters covering all known human genes was produced by taking the KnownGenes table from the UCSC genome browser and keeping only the transcript with the longest 5′-UTR if a gene has multiple transcripts with the same translation start site. Fifteen 50-mer probes were designed for each promoter.

To identify the promoters bound by YY1, or ChIP hits, the hybridization signal of the dye swapping experiments was averaged and associated with the genomic coordinates of the probes. A signal cutoff was set to the mean plus 2.5 standard deviations. A hit was called for a genomic region covered by at least five probes above this cutoff, allowing gaps smaller than 200 bp between consecutive probes. Genes immediately adjacent to the ChIP hits were identified by finding the TSS closest to the boundaries of hits. These genes were later used to analyze 5′-UTR length and enrichment of GO terms.

### Positional specificity of ChIP signal

The raw data of a ChIP-chip experiment was formatted so that each probe was represented as its mapped genomic coordinates associated with a hybridization signal level. Given a certain set of anchor coordinates, such as those of TSSs, we calculated the relative genomic distance between each probe and its nearest anchor. Then, the signal associated with each probe was mapped to the specific distance from the anchor, and signals at the distance were averaged across all anchors. The signal was further smoothened by averaging over a moving window of 100 data points. Finally, the averaged and smoothened signals were plotted against the distance to the anchors.

### Calculation of 5′-UTR length

5′-UTR length was calculated as the distance between the translation start site and the transcriptional start site. For the genes in the 723 high-quality DBTSS promoter set, the coordinates of translation start sites were first obtained from RefSeq and the coordinates of the transcriptional start sites were obtained from DBTSS. For the genes with YY1 ChIP hits and randomly selected sequences, the coordinates of transcription start sites and translation start sites were both obtained from RefSeq.

## Acknowledgments

We thank Bing Ren for providing us with the YY1 ChIP-chip data set prior to publication. We thank Tom Tullius for letting us use his radioactive facility. We thank The ENCODE Project Consortium for making their data publicly available and Transcriptional Regulation analysis group for providing ChIP-chip and ChIP-sequencing data sets. We thank the *Genome Research* reviewers for their insightful comments. We thank Jesscia Marie Barros and Enoch Huang for proofreading the manuscript. This work was funded by the ENCODE Consortium grant R01HG03110 from NHGRI, NIH to Z.W.

## References

Athanikar, J.N., Badge, R.M., and Moran, J.V. 2004. A YY1-binding site is required for accurate human LINE-1 transcription initiation. *Nucleic Acids Res.* **32:** 3846–3855.

Bailey, T.L. and Elkan, C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2:** 28–36.

Bailey, T.L., Baker, M.E., and Elkan, C.P. 1997. An artificial intelligence approach to motif discovery in protein sequences: Application to steriod dehydrogenases. *J. Steroid Biochem. Mol. Biol.* **62:** 29–44.

Becker, K.G., Swergold, G.D., Ozato, K., and Thayer, R.E. 1993. Binding of the ubiquitous nuclear transcription factor YY1 to a *cis* regulatory sequence in the human LINE-1 transposable element. *Hum. Mol. Genet.* **2:** 1697–1702.

Beissbarth, T. and Speed, T.P. 2004. GOstat: Find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* **20:** 1464–1465.

Blanchard, J.L. and Lynch, M. 2000. Organellar genes: Why do they end up in the nucleus? *Trends Genet.* **16:** 315–320.

Burke, T.W. and Kadonaga, J.T. 1996. *Drosophila* TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. *Genes & Dev.* **10:** 711–724.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C.A., Taylor, M.S., Engstrom, P.G., Frith, M.C., et al. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* **38:** 626–635.

Cole, E.G. and Gaston, K. 1997. A functional YY1 binding site is necessary and sufficient to activate Surf-1 promoter activity in response to serum growth factors. *Nucleic Acids Res.* **25:** 3705–3711.

The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* (in press).

FitzGerald, P.C., Shlyakhtenko, A., Mir, A.A., and Vinson, C. 2004. Clustering of DNA sequences in human promoters. *Genome Res.* **14:** 1562–1574.

Hariharan, N., Kelley, D.E., and Perry, R.P. 1991. Delta, a transcription factor that binds to downstream elements in several polymerase II promoters, is a functionally versatile zinc finger protein. *Proc. Natl. Acad. Sci.* **88:** 9799–9803.

Haverty, P.M., Frith, M.C., and Weng, Z. 2004a. CARRIE Web service: Automated transcriptional regulatory network inference and interactive analysis. *Nucleic Acids Res.* **32:** W213–W216.

Haverty, P.M., Hansen, U., and Weng, Z. 2004b. Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification. *Nucleic Acids Res.* **32:** 179–188.

Huber, R., Schlessinger, D., and Pilia, G. 1998. Multiple Sp1 sites efficiently drive transcription of the TATA-less promoter of the human glypican 3 (GPC3) gene. *Gene* **214:** 35–44.

Javahery, R., Khachi, A., Lo, K., Zenzie-Gregory, B., and Smale, S.T. 1994. DNA sequence requirements for transcriptional initiator activity in mammalian cells. *Mol. Cell. Biol.* **14:** 116–127.

Kozak, M. 1984. Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. *Nucleic Acids Res.*

**12:** 857–872.
Kozak, M. 1987a. An analysis of 5′-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res.* **15:** 8125–8148.
Kozak, M. 1987b. At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells. *J. Mol. Biol.* **196:** 947–950.
Lavie, L., Maldener, E., Brouha, B., Meese, E.U., and Mayer, J. 2004. The human L1 promoter: Variable transcription initiation sites and a major impact of upstream flanking sequence on promoter activity. *Genome Res.* **14:** 2253–2260.
Li, W.W., Hsiung, Y., Wong, V., Galvin, K., Zhou, Y., Shi, Y., and Lee, A.S. 1997. Suppression of grp78 core promoter element-mediated stress induction by the dbpA and dbpB (YB-1) cold shock domain proteins. *Mol. Cell. Biol.* **17:** 61–68.
Lim, C.Y., Santoso, B., Boulay, T., Dong, E., Ohler, U., and Kadonaga, J.T. 2004. The MTE, a new core promoter element for transcription by RNA polymerase II. *Genes & Dev.* **18:** 1606–1617.
Mantovani, R. 1998. A survey of 178 NF-Y binding CCAAT boxes. *Nucleic Acids Res.* **26:** 1135–1143.
Ohler, U., Liao, G.C., Niemann, H., and Rubin, G.M. 2002. Computational analysis of core promoters in the *Drosophila* genome. *Genome Biol.* **3:** RESEARCH0087.
Riggs, K.J., Saleque, S., Wong, K.K., Merrell, K.T., Lee, J.S., Shi, Y., and Calame, K. 1993. Yin-yang 1 activates the c-myc promoter. *Mol. Cell. Biol.* **13:** 7487–7495.
Safrany, G. and Perry, R.P. 1995. The relative contributions of various transcription factors to the overall promoter strength of the mouse ribosomal protein L30 gene. *Eur. J. Biochem.* **230:** 1066–1072.
Shi, Y., Seto, E., Chang, L.S., and Shenk, T. 1991. Transcriptional repression by YY1, a human GLI-Kruppel-related protein, and relief of repression by adenovirus E1A protein. *Cell* **67:** 377–388.
Shi, Y., Lee, J.S., and Galvin, K.M. 1997. Everything you have ever wanted to know about Yin Yang 1. *Biochim. Biophys. Acta* **1332:** F49–F66.
Shrivastava, A. and Calame, K. 1994. An analysis of genes regulated by the multi-functional transcriptional regulator Yin Yang-1. *Nucleic Acids Res.* **22:** 5151–5155.
Smale, S.T. 1997. Transcription initiation from TATA-less promoters within eukaryotic protein-coding genes. *Biochim. Biophys. Acta* **1351:** 73–88.
Smale, S.T. and Baltimore, D. 1989. The "initiator" as a transcription control element. *Cell* **57:** 103–113.
Smale, S.T. and Kadonaga, J.T. 2003. The RNA polymerase II core promoter. *Annu. Rev. Biochem.* **72:** 449–479.
Smit, A.F. 1996. The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Dev.* **6:** 743–748.
Smith, E., Meyerrose, T.E., Kohler, T., Namdar-Attar, M., Bab, N., Lahat, O., Noh, T., Li, J., Karaman, M.W., Hacia, J.G., et al. 2005. Leaky ribosomal scanning in mammalian genomes: Significance of histone H4 alternative translation in vivo. *Nucleic Acids Res.* **33:** 1298–1308.
Srinivasan, L. and Atchison, M.L. 2004. YY1 DNA binding and PcG recruitment requires CtBP. *Genes & Dev.* **18:** 2596–2601.
Thomas, M.J. and Seto, E. 1999. Unlocking the mechanisms of transcription factor YY1: Are chromatin modifying enzymes the key? *Gene* **236:** 197–208.
Wei, C.L., Wu, Q., Vega, V.B., Chiu, K.P., Ng, P., Zhang, T., Shahab, A., Yong, H.C., Fu, Y., Weng, Z., et al. 2006. A global map of p53 transcription-factor binding sites in the human genome. *Cell* **124:** 207–219.

*Received July 13, 2006; accepted in revised form January 9, 2007.*