Published online 11 November 2010 *Nucleic Acids Research, 2011, Vol. 39, No. 6* ***2175–2187***
doi:10.1093/nar/gkq1035

Genome-wide detection of a TFIID localization element from an initial human disease mutation

Mary Q. Yang[1], Karina Laflamme[2], Valer Gotea[1], Clinton H. Joiner[3], Nancy E. Seidel[2], Clara Wong[4], Hanna M. Petrykowska[1], Jens Lichtenberg[5], Stephen Lee[6], Lonnie Welch[5,7,8], Patrick G. Gallagher[4], David M. Bodine[2] and Laura Elnitski[1,*]

[1]Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Rockville MD 20852, [2]Genetics and Molecular Biology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda MD 20892, [3]Cincinnati Children's Hospital Research Foundation, Division of Hematology/Oncology Cincinnati OH 45229, [4]Department of Pediatrics, Yale University School of Medicine, New Haven CT 06520, [5]School of EECS, Ohio University, Athens, Ohio 45701, USA, [6]Department of Statistics, University of Idaho, Moscow, Idaho 83844, USA, [7]Biomedical Engineering Program and [8]Molecular and Cellular Biology Program, Ohio University, Athens, OH 45701, USA

Received February 2, 2010; Revised October 5, 2010; Accepted October 10, 2010

ABSTRACT

Eukaryotic core promoters are often characterized by the presence of consensus motifs such as the TATA box or initiator elements, which attract and direct the transcriptional machinery to the transcription start site. However, many human promoters have none of the known core promoter motifs, suggesting that undiscovered promoter motifs exist in the genome. We previously identified a mutation in the human Ankyrin-1 (*ANK*-1) promoter that causes the disease ankyrin-deficient Hereditary Spherocytosis (HS). Although the *ANK*-1 promoter is CpG rich, no discernable basal promoter elements had been identified. We showed that the HS mutation disrupted the binding of the transcription factor TFIID, the major component of the pre-initiation complex. We hypothesized that the mutation identified a candidate promoter element with a more widespread role in gene regulation. We examined 17 181 human promoters for the experimentally validated binding site, called the TFIID localization sequence (DLS) and found three times as many promoters containing DLS than TATA motifs. Mutational analyses of DLS sequences confirmed their functional significance, as did the addition of a DLS site to a minimal Sp1 promoter. Our results demonstrate that novel promoter elements can be identified on a genome-wide scale through observations of regulatory disruptions that cause human disease.

INTRODUCTION

Eukaryotic protein-coding genes are transcribed by RNA polymerase (RNA Pol II) and are referred to as class II genes. RNA Pol II interacts directly with the core promoter sequence, which is defined as the minimal region required to direct low levels of accurately initiated transcription *in vitro* (1). Core promoters encompass the transcription start site (TSS) and flanking sequences from −50 to +50 (2–6). As the fundamental step of transcription initiation, core promoters recruit the basal transcription machinery [or pre-initiation complex (PIC)]. This large, multi-component regulatory complex anchors to the DNA of the core promoter through sequence-specific DNA contacts made by one of its numerous subunits.

A handful of core-promoter elements have been proven to act as docking sites for sequence-specific contacts. The most well known elements are TATA box, initiator (Inr), downstream promoter element (DPE) and TFIIB recognition element (BRE) (3). Each of these is characterized by a preferred binding sequence in the DNA, often represented as a position weight matrix (PWM) that captures the frequency of nucleotide combinations and allowable variation. The majority of mammalian core promoters contain CpG-islands (7,8). Although not considered a basal promoter element, CpG-islands are recognized as important features of promoters, which

*To whom correspondence should be addressed. Tel: +1 301 451 0265; Fax: +1 301 435 6170; Email: elnitski@mail.nih.gov

maintain open chromatin and protect against the repressive effects of DNA methylation. Basal promoter elements can be embedded in CpG-islands, however, to date, only Sp1 binding sites have been shown to have a large occurrence within CpG-islands. Additional sequence-specific elements are plausible, although the di-nucleotide composition of CpG rich regions complicates detection.

Although there is no universal core promoter element, recurring use of several common recognition elements is a general theme. Moreover, combinations of core-promoter elements often add a level of refinement necessary to modulate expression through additive, synergistic or even repressive effects (9). These include TATA box and Inr (10,11), DPE and Inr (12,13) and BRE and TATA box (14,15). More recently BRE and Inr, TATA and DPE and BRE and DPE have been identified as functional combinations that provide additional regulatory refinement, albeit in lower proportions of all promoters (16). Many promoters lack a recognizable basal promoter element, fueling speculation that novel functional promoter elements will continue to be discovered (9–17).

Core promoters are classified by the distribution of their transcription initiation sites. Approximately one-third of core promoters in vertebrates are narrow-peaked or focused, with a single or closely clustered TSS location (18–20). The remaining vertebrate core promoters are known as broad or dispersed, with multiple TSSs distributed over a 50–100 bp region. These often correspond to CpG-island promoters (21,22) where the Sp1 protein may direct the basal machinery to form a PIC (3). Multimers of two or more adjacent Sp1-binding sites are recognized to synergistically activate transcription (23). Sp1 sites are often found in a region 40–80 bp upstream of the TSS, which conforms to the need for proximity to the TSS in PIC recruitment, yet reveals a broad range of allowable positions rather than a single functional location.

We have previously reported that the human ankyrin-1 (*ANK*-1) promoter has a high G + C content (77%) and a broad initiation pattern containing multiple, closely spaced TSSs (24). Furthermore, the gene utilizes several alternative promoters to confer tissue specificity to its transcript isoforms (25). Despite the appearance of a CpG-rich housekeeping-like promoter, the *ANK*-1 promoter is erythroid-specific and contains functional Sp1 binding sites, but lacks TATA, Inr, CCAAT and other core promoter sequences. Nevertheless, a novel functional element with an indispensible role in the erythroid *ANK*-1 promoter was identified at the site of a deletion mutation in a patient with ankyrin-deficient Hereditary Spherocytosis (HS) (26). The omission of a TG di-nucleotide from the promoter sequence decreased both binding of the TATA box binding protein (TBP) subunit of TFIID and transcription of *ANK*-1 mRNA. As a result, the choice of initiation site utilized in transgenic mice and cell-free transcription assays became altered (27). Further experimental analyses of the DNA at the deletion site identified a consensus motif that bound TFIID and was essential for transcription *in vitro* and *in vivo* (24). One particular version of the TFIID localization sequence (DLS) motif conferred a 5-fold increase over the wild-type DLS sequence when assayed in transgenic mice, illustrating the potential for modulation of regulatory activity through site-specific variation in this binding motif. In this report, we describe the regulatory landscape of the human genome with respect to the DLS motif and test whether DLS sites are necessary to support transcription. We conclude that the frequency with which DLS occurs in human promoter sequences coupled with the functional impact of mutations to these motifs provides strong evidence that the DLS should be considered a novel TFIID localization element in promoters.

MATERIALS AND METHODS

Transient transfection analyses

The tissue culture cell line K562 (erythroid) was maintained in Improved Eagle's minimal essential medium (Invitrogen, Gaithersburg, MD, USA), containing 10% fetal calf serum (Hyclone, Logan, UT, USA). Four DLS promoter sequences (human *FBXO18* and *KCC3* or mouse *Tesc* and *Klf3*) were subcloned into a promoterless firefly luciferase reporter plasmid and sequence verified. Mutations were introduced by replacing the DLS sequence with ATTAACAGA. Sp1 sites or DLS sites were designed using synthetic oligonucleotides and cloned into the promoter-less vector. K562 cells (10^7) were transfected by electroporation with a single pulse of 300 V at 950 μF with 20 μg of test plasmid and 0.5 μg of pRL-SV40, a mammalian reporter plasmid expressing *Renilla* luciferase driven by the Simian Virus 40 (SV40) early gene promoter (Promega, Madison, WI, USA). Forty-eight hours after transfection, cells were harvested, lysed and the ratio of firefly luciferase (test) and *Renilla* luciferase activity (control) were determined using a Fluoroskan Ascent FL (Thermo, Gaithersburg, MD, USA). All assays were performed in triplicate.

Computational analyses

TSSs were collected from the v5 dbTSS repository (28). In total, 17 181 TSSs were recorded from clustered CAGE tag data annotated on the human genome sequence assembly hg17 and converted to hg18, using the 'liftOver' tool of the UCSC Human Genome Browser. Unclustered CAGE tags were collected from the v6 release of dbTSS. The genomic position of each TSS was used to calculate the coordinates of the core promoter region from 100 bp upstream of each TSS through 40 bp downstream. The positions of DNA motifs within the sequences were mapped using pattern matching with regular expressions in Perl. These motifs included TATA, Inr, CC AAT, DPE, BRE and DLS. All motifs except DLS were counted only in the position relevant to their known functional activity. For instance, TATA motifs were directional, located within a window 30 bp upstream of the TSS and had to match the 7 bp motif recorded below. Combinations of motifs or the absence of any features in a promoter were determined through intersecting the genomic coordinates. Binding sites were mapped within the region of the expected functional location allowing some padding for small windows to accommodate

imprecise TSS mapping, for example TATA motifs (−49, −20), DPE from (24, 34), Inr at (−15, 15), BRE at (−44, −18), CCAAT at (−108, 9) and Sp1 (−80, −40). DLS motifs were initially assessed in the range from (−100 to +40) and later expanded to (−550 to +100). Motifs that overlapped the boundaries of the window were included. Subsequent analyses examined positions −550 to +1000. The following consensus sequences were used:

TATA-TATA(A|T)A(A|G|T); Inr-(C|T)(C|T)A(A|C|G|T)(A|T)(C|T)(C|T); CCAAT-(A|G)(A|G)CCAAT(A|C|G)(A|G); DPE-(A|G|T)(C|G)(A|T)(C|T)(A|C|G)(C|T); BRE-(G|C)(G|C)(G|A)CGCC; SP1-G(A|G)GGC(A|G)GGG(A|T); DLS-(T|G)(C|G)(C|G)GGNGAG.

Genomic positions of CpG-islands were obtained from the Human UCSC Genome Browser. A chi-square test for independence was the statistical test for the initial randomized sequences. As an additional filter, 63 promoter sequences were identified to overlap a 5′-splice-donor site (3.5%), which carries a partial match to the 'GTGAG' nucleotides. To assess whether the DLS motif occurrence was significantly larger than what was expected by chance given the mono- or di-nucleotide frequencies, we constructed 1000 sets of 17 181 sequences, all with the same length as the sequences in the initial set of promoters by randomly sampling 1 or 2 nt at a time from the sequences of core promoters. The expected number of sequences containing DLS was also computed theoretically using the mono- and di-nucleotide frequencies in core promoter sequences and the approach described in (29). To further test whether the enrichment of DLS motif observed in core promoter regions was simply due to the presence of CpG-islands, we created 10 000 additional sets of CpG-island regions located outside of core promoters through random sampling; each set containing the same number of sequences as in the promoter CpG-island set. The 10 000 sets of sequences were used to construct a distribution of the expected occurrence of the DLS motif in CpG island regions, to which we compared the occurrence of the DLS motif in CpG islands in core promoter regions. To minimize the possibility that CpG island sequences were located in core promoter regions, we excluded from the randomization process all CpG island segments that overlap with −100 to +40 windows around all TSSs defined in dbTSS v5, v6, (ftp://ftp.hgc.jp/pub/hgc/db/dbtss/dbtss_ver6/hspromoter.tab.gz) as well as around TSSs of known transcripts as annotated in RefSeq and known gene sets.

The Markov chain model calculation was performed with WordSeeker, a software suite for discovering patterns and features in genomic sequences (http://wordseeker.org/). A radix tree data structure enumerated substrings of 9-mer present in the promoter sequences. The expected number of occurrences E of each word w was calculated as as $E = (n - |w| + 1) * p_w$, where n is the total number of 9-mers present in the promoter sequences masked for repetitive elements $|w|$ is the length of word w and p_w is the probability of word observing word w [calculated by using an order 1 Markov model (30)]. The P-value of each O-mer (DNA word of length O) was calculated using a binomial test of statistical significance.

Analyses of evolutionary conservation in DLS motifs utilized five species (human, chimp, mouse, rat and dog) multiZ alignments obtained from the UCSC Human Genome Browser. DLS motifs were identified in the human sequence and assessed for an identical match at the orthologous positions in other species.

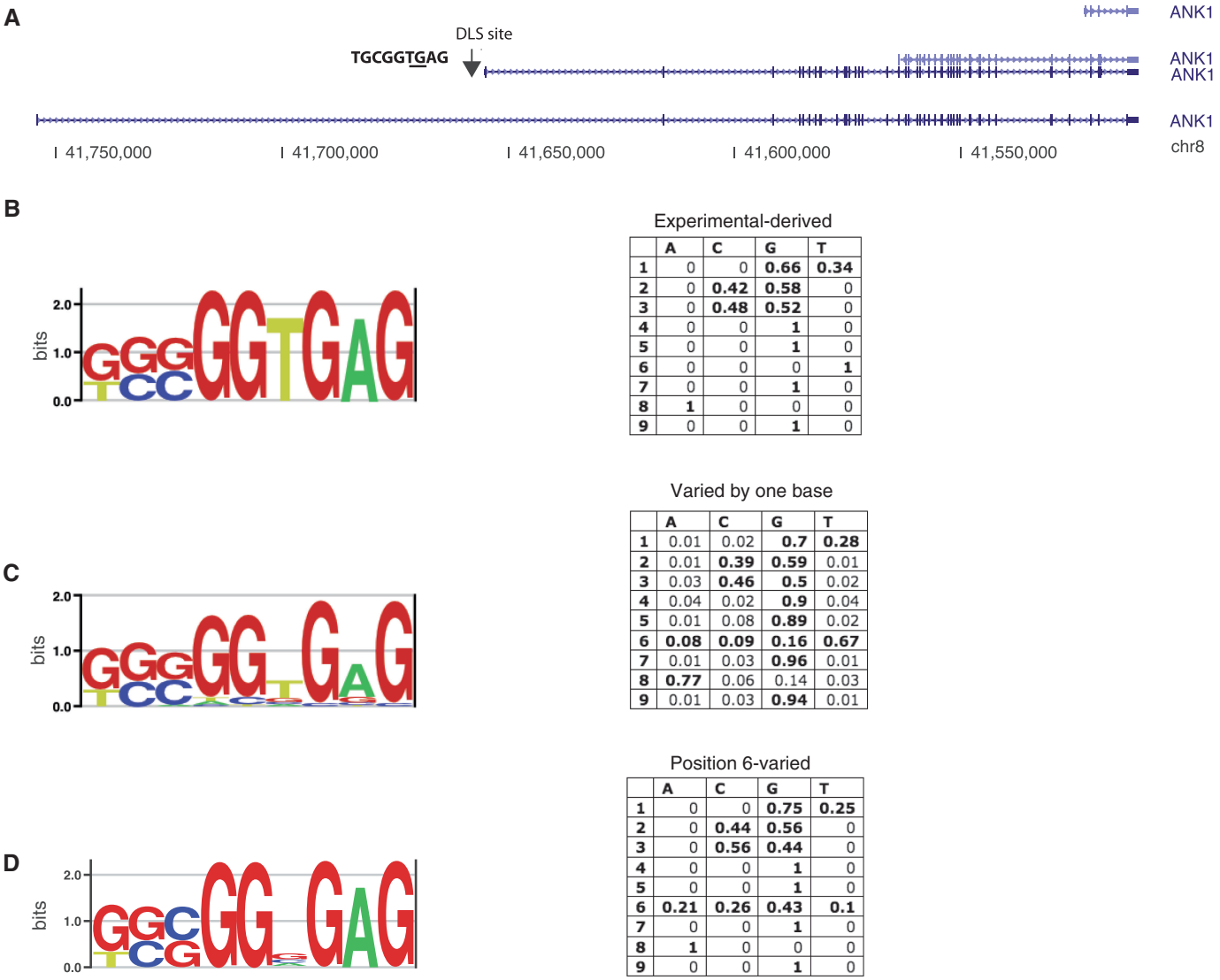
Ultraconserved elements were from a collection of 481 segments longer than 200 bp with absolute conservation (100% identity with no insertions or deletions) between orthologous regions of the human, rat and mouse genomes (31). A second collection of ultraconserved elements was from mammalian promoters (32), containing 2827 regions in 5 kb upstream regions of 1268 human protein-coding genes with 98% identity for at least 30 bp in seven mammalian species. Genomic coordinates were used to intersect the locations of the DLS motifs and these data. Gene ontology (GO) terms were examined using the GOSTAT server with Benjamini correction, minimum sub-GO length of 3 (http://gostat.wehi.edu.au/).

RNA Pol II, DNaseI HS and formaldehyde assisted isolation of regulatory elements (FAIRE) data were downloaded from the UCSC Human Table Browser under the heading of ENCODE Open Chromatin, Duke/UNC/UT (33,34). Sites were collated and compared with DLS or TATA motifs based on genomic coordinates. UCSC data labeled 'ChIP-seq Pol2 peaks', 'DNaseI-Seq peaks' and 'FAIRE-Seq peaks' were compiled from HeLa, K562 and GM12878 cell data.

RESULTS

Genome-wide prediction of functional DLS sites

The TFIID localization sequence (DLS) was identified through an analysis of a dinucleotide deletion in the *ANK-1* promoter of a patient with HS (Figure 1A; 32). The role of the DLS in the *ANK-1* promoter was functionally confirmed in our recent work by *in vitro* foot-printing, mutational analyses and expression analyses in transgenic mice (24). Initially defined from a library of ~16 000 cloned synthetic oligos, the functional sequence was assessed for an ability to support transcription from the *ANK-1* DLS site through *in vitro* run-on assays. A 9-bp IUPAC consensus motif (K)(S)(S)GGTG AG was identified from four recurrent sequences (Figure 1B). To assess the prevalence of the DLS motif in promoters of the human genome we screened an *in silico* library of 17 181 predicted promoter regions that we created from the dbTSS CAGE data (28; http://dbtss.hgc.jp/). From these data we extracted minimal promoter sequences, defined as the regions from −100 to +40 around each TSS and determined the percent of sequences matching the IUPAC DLS consensus. A majority of these 17 181 promoters overlapped CpG islands (60%), as is consistent with previous publications describing human promoter sequences (35). For comparison, sets of control sequences were examined, offering increasing stringency for background composition. The DNA sequences in the controls were matched to have the same



	A	C	G	T
1	0	0	**0.66**	**0.34**
2	0	**0.42**	**0.58**	0
3	0	**0.48**	**0.52**	0
4	0	0	**1**	0
5	0	0	**1**	0
6	0	0	0	**1**
7	0	0	**1**	0
8	**1**	0	0	0
9	0	0	**1**	0

	A	C	G	T
1	0.01	0.02	**0.7**	**0.28**
2	0.01	**0.39**	**0.59**	0.01
3	0.03	**0.46**	**0.5**	0.02
4	0.04	0.02	**0.9**	0.04
5	0.01	0.08	**0.89**	0.02
6	**0.08**	**0.09**	**0.16**	**0.67**
7	0.01	0.03	**0.96**	0.01
8	**0.77**	0.06	0.14	0.03
9	0.01	0.03	**0.94**	0.01

	A	C	G	T
1	0	0	**0.75**	**0.25**
2	0	**0.44**	**0.56**	0
3	0	**0.56**	**0.44**	0
4	0	0	**1**	0
5	0	0	**1**	0
6	**0.21**	**0.26**	**0.43**	**0.1**
7	0	0	**1**	0
8	**1**	0	0	0
9	0	0	**1**	0

Figure 1. Development of a DLS motif pattern for promoter analyses. An illustration of the *ANK-1* locus showing alternative promoters (**A**). The arrow at the erythroid-specific promoter marks the position and 9 bp sequence of the DLS motif. The 'TG' dinucleotide deleted in a patient with ankyrin-deficient HS is underlined. (**B**) A sequence logo illustrates the nucleotide composition of the four DLS sites that functioned in an experimental analysis of the *ANK-1* promoter. Adjacent to the logo is the PFM generated from all motifs identified in the promoter sequences using the IUPAC consensus motif. The consensus string was expanded by one mismatch at any position (**C**) and then held constant to incorporate any nucleotide at Position 6 (**D**). The most frequently encountered nucleotides at each position are shown in bold.

number as the 17 181 promoter sequences. The controls comprised randomly selected regions of the human genome and background models based on 1 and 2-bp sampling of the promoter sequences to maintain the CpG content while scrambling the higher-order combinations. A count of the experimentally derived 9-bp DLS consensus (24) in the promoter sequences identified 1.3% of the promoters, whereas a search of 17 181 randomly sampled sequences from the genome identified only 0.2% of the total, representing a 6.5-fold enrichment ($P < 1.12 \times 10^{-9}$). The significant enrichment of the DLS motif in promoter sequences compared with control data implicated the DLS in widespread transcriptional regulatory processes in the genome, but did not address the many complexities associated with motif searches in promoter sequences. Moreover, this consensus motif was based on experimental data derived from a small number of examples (24) in a stringent *in vitro* assay system, which could underestimate the nucleotide diversity allowable at promoters other than *ANK*-1.

We addressed the spectrum of variation that could be accommodated by DLS sites across the genome by allowing single base mutations to the IUPAC consensus and evaluating changes in the position frequency matrix (PFM) (Figure 1B). This search strategy found 3835/17 181 examples in the promoter sequences (22.3%) compared with 5.9% of the random sequenes ($P < 2.2 \times 10^{-16}$) and generated a new PFM that was significantly

different only at Position 6 (Figure 1C). Because Positions 6 and 7 were implicated as necessary in the original HS mutation, they were held constant as 'TG' in the experimental assessment of LaFlamme *et al.* (24). However, our computational results clearly indicated that Position 6 contained T in 67% of sequences and allowed G, A or C in 16%, 9% and 8% of the sequences, respectively. The relative entropy levels depicted by the sequence logo in Figure 1C illustrate the variability at Position 6. In contrast, the neighboring Position 7 remained the same as the reference sequence in the human *ANK-1* promoter (i.e. 'G') in 96% of the sequences (see PFM in Figure 1C). We concluded that the nucleotide identity at Position 6 was interchangeable, but not dispensable, as shown by previous clinical and experimental analyses of deletions at this site in human cells and transgenic mice (24–26).

Based on the expanded search for DLS motifs in the promoter sequences, a new IUPAC consensus sequence was derived wherein Position 6 was allowed to include any nucleotide, (K)(S)(S)GGNGAG (Figure 1D). The new motif was present in 9.4% of human promoter sequences (1621), compared with 1% of control sequences (170) that were sampled 1000 times from non-core promoter regions of the genome (9.1-fold difference, $P < 1 \times 10^{-3}$) (Figure 2A).

Since the expansion of the motif (based on a 1-bp mismatch at any position from the consensus) showed a significant change for only one position, we repeated the search allowing mismatches at any two positions. The PFM showed a decrease at all consensus positions from the previous analysis and an increase in all non-consensus positions (Supplementary Figure S1). A search for this expanded pattern revealed a significant loss of motif specificity, as concluded from the fact that two matches to the pattern were often predicted in each of the 17 181 promoter sequences. All subsequent analyses used the (K)(S)(S)GGNGAG motif.

Additional tests were implemented to rule out the influence of high G+C content in core promoter regions as a cause of the observed DLS enrichment. The mono-nucleotide and di-nucleotide backgrounds of the promoter sequences were modeled by 1- or 2-bp sampling of the sequences to produce 1000 independent sets of 17 181 control sequences (Figure 2A). Compared with the distribution of sequences sampled from random genomic DNA, the mono- and di-nucleotide sampling of core promoter sequences showed an increase in the number of DLS sequences. However, the observed number of DLS motifs identified in the promoter sequences was larger and remained statistically significant relative to the distributions in the background models ($P < 1 \times 10^{-3}$ for each comparison, Figure 2A), indicating that high G+C background was a contributing factor, but not the sole cause of DLS enrichment in promoter sequences. We also compared actual DLS occurrences in core promoters with theoretical expectations using background frequencies of mono- and di-nucleotides in core promoters ($P < 9.3 \times 10^{-79}$ and $P < 4.7 \times 10^{-24}$, respectively, using the binomial test). The simulated results showed precise agreement with the theoretical estimates [Figure 2A see (29) for the method].

To test whether the enrichment of the DLS in core promoters is due to the high frequency of CpG-island regions, we compared the DLS occurrence in CpG-islands that fell within or outside of core promoters ('Materials and Methods' section). In total, 9888 such core promoter regions had 1114 DLS motifs (11%), whereas a 33% decrease was recorded for regions outside of core promoters ($P < 1 \times 10^{-4}$; Figure 2B). These data confirmed that the DLS had a significant presence in the TSS-proximal regions compared with distal regions.

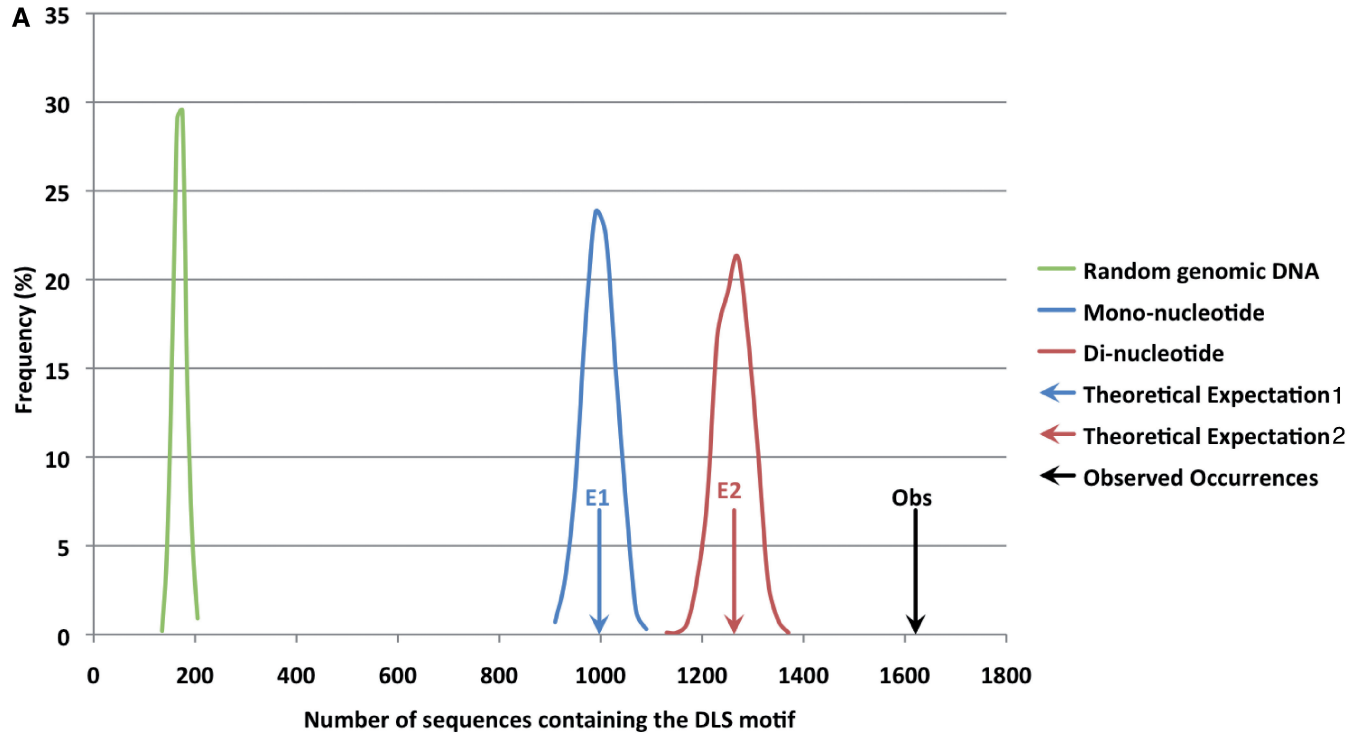
Enrichment of DLS strings

Preferred strings of the DLS were examined using a Markov chain. In total, 12 strings of the DLS motif were overrepresented (Table 1, 'Materials and Methods' section). The string with the lowest *P*-value for significance was GGCGGGGAG ($P < 7.73 \times 10^{-11}$). Our previous work showed that small changes in the DLS consensus sequence caused dramatic changes in the transcript expression levels (24), thus these varied DLS motifs are likely to contribute individualized activity.

DLS positions relative to TSSs

To address whether the DLS motif occurred beyond the proximal promoter region, we mapped the motif from positions −550 to +1000 and found that DLS occurrences peaked near the TSS and returned to baseline levels both upstream and downstream (Figure 3A). The random sequences showed continuous low-level signal for DLS motifs, but no distinctive peaks (Figure 3B, red line). DLS occurrences increased markedly in the proximal promoter region from positions −100 to +40 compared with distal positions −550 to −100. The increased occurrences continued downstream of the TSS and well beyond position +40, including a second peak at position +250. Like the DLS motif, CCAAT and Sp1 profiles showed broad curves of enrichment starting at position −200 and peaking further upstream of the TSS than TATA (Figure 3B–E), although the broad, upward sloping profile of DLS was wider than any other motif. We hypothesized that the breadth of the DLS profile could result from the dispersed nature of TSSs in CpG-rich promoters creating breadth to the distance measurements. To address this idea, each TSS in the 17 181 sequences [from dbTSS v5 (28, 36)] was expanded to include all surrounding CAGE tags (i.e. unclustered as 101436 individual data points) from the more recent dbTSS release (v6). We found that the breadth of transcriptional initiation sites remained narrower than the DLS enrichment profile (gray versus blue lines; Figure 3A). These data indicated that dispersed TSSs were not creating the broad DLS profile. Furthermore, no significant peaks appeared in the TSS CAGE data to support the hypothesis of alternative downstream promoters adding breadth to the DLS profile.

Although most promoter elements are reported within a narrow region around the TSS position (20), we mapped several motifs in a 1550 bp window. Another core promoter element showing a broad zone of enrichment similar to DLS was BRE (Figure 4), which functions both upstream (BRE-u) and downstream (BRE-d) of the



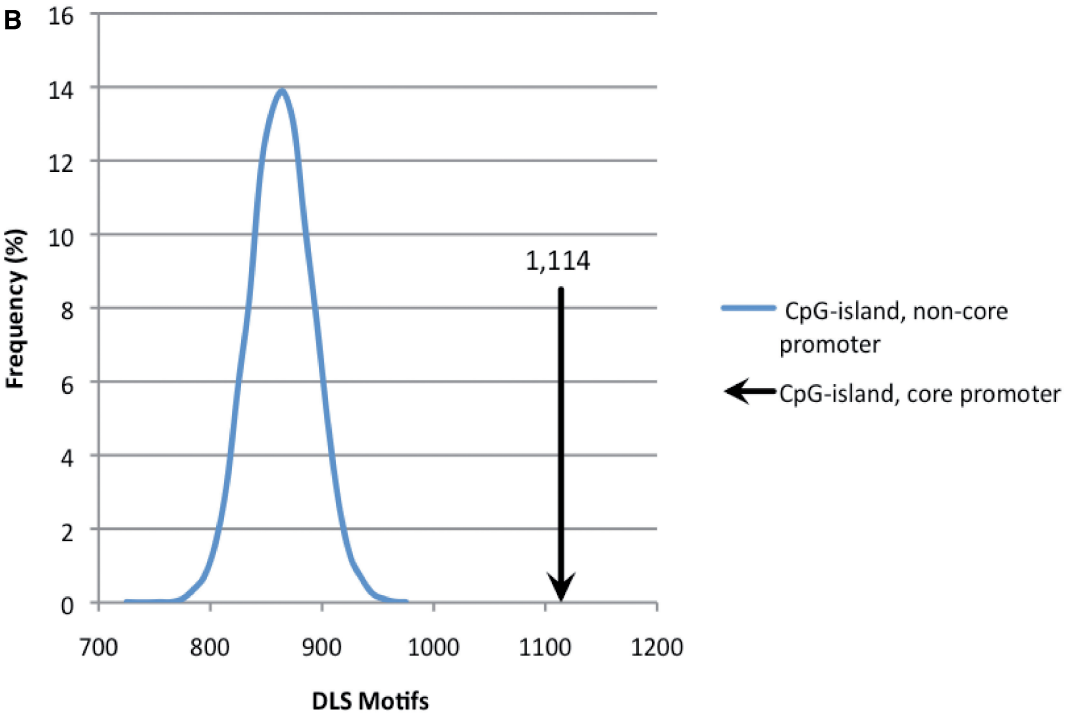


Figure 2. (**A**) The occurrence of the DLS motif in genomic sequences. Comparisons between core promoters and control sequences were conducted by randomly sampling 1 bp (blue) and 2 bp (red) from core promoter regions to generate 1000 replicate sets of scrambled sequences with similar CpG content. Additionally, 156 bp sequences from other genomic regions (green) were assessed for DLS occurrences in 1000 replicates. The black arrow corresponds to the 1621 sequences that contain at least one DLS motif. The arrows marked by E1 and E2 indicate expected values obtained theoretically based on the frequency of mono- and di-nucleotides, respectively, as described in (29). (**B**) Occurrence of the DLS motif in CpG islands. CpG islands were classified into core promoter regions and non-core promoter regions. DLS occurrences in the core promoter group (black arrow) were compared the non-core promoter group, for which the distribution was measured in 10 000 sets of randomly selected sequences.

Table 1. Statistically significant substrings of the DLS motif

Word	O	E	O/E	*P*-value (<)
GGCGGGGAG	158	50.68	3.11	$<7.73 \times 10^{-11}$
GGCGGCGAG	79	34.66	2.27	$<7.73 \times 10^{-11}$
GCGGGGGAG	99	50.68	1.95	$<1.27 \times 10^{-09}$
GGCGGAGAG	68	32.11	2.11	$<2.39 \times 10^{-08}$
GCGGGCGAG	65	34.66	1.87	$<2.74 \times 10^{-06}$
GCCGGGGAG	76	48.64	1.56	$<1.71 \times 10^{-04}$
GGCGGTGAG	34	21.09	1.6	$<5.87 \times 10^{-03}$
GCCGGCGAG	48	33.26	1.44	$<9.53 \times 10^{-03}$
GCCGGTGAG	30	20.24	1.48	$<2.50 \times 10^{-02}$
TGGGGTGAG	27	18.25	1.47	$<3.27 \times 10^{-02}$
TGGGGAGAG	38	27.79	1.36	$<3.78 \times 10^{-02}$
TCCGGGGAG	37	27.67	1.33	$<5.18 \times 10^{-02}$

The observed number of occurrences (O)/expected number of occurrences (E) is labeled O/E.

TATA box (33). In our analysis BRE showed a secondary peak near position +250. Both DLS and BRE are enriched in the sequences where G+C content is high. To ensure DLS and BRE motifs were not mapping to either the same or overlapping positions in promoters, motif redundancy was measured for DLS and TATA, CCAAT, INR, DPE and BRE. No exact matches were possible from the consensus motifs and <16% of the DLS motifs showed a partial overlap with another motif (Table 2). Thus the co-occurring peaks at the +250 position suggest that DLS and BRE may be functionally relevant even when located far downstream of the TSS.

In this analysis the experimental DLS motif was mapped only in the direction of transcription, consistent with both the approach for mapping TATA motifs and with the orientation of the original DLS in the *ANK-1* promoter. We also examined the occurrence of the DLS motif on the reverse strand of the promoter sequences. In contrast to the elevated presence of the DLS on the forward strand continuing to position +250, we found a sharp decrease in the reverse DLS motifs mapping downstream of the TSS (Supplementary Figure S2).

Functional analysis of DLS motifs

To address whether the predicted DLS motifs represented functional elements in gene expression assays, four DLS promoters were identified and examined in a luciferase reporter assay. Each contained a single DLS (human *FBXO18* and human *KCC3* or mouse *Tesc* and mouse *Klf3*) where the sequence was either an exact match or a single mismatch from the IUPAC consensus (Table 3). All 4 wild-type promoters showed functional activity in luciferase-reporter assays in K562 cells (Figure 5). The DLS for each promoter was positioned at a different distance from the annotated TSS in the CAGE data (Table 3). However, all four genes had evidence for multiple closely spaced initiation sites, which was consistent with CpG-rich promoters and was verified by EST data in the UCSC Genome Browser (not shown). The DLS in each of these promoters was replaced by ATTA ACAGA. The altered promoters were cloned into luciferase expression vectors and transiently transfected in K562 cells. Similar to the loss of function observed from the *ANK*-1 promoter deletion, mutagenesis of these DLS sequences significantly reduced luciferase expression by 2- to 10-fold (Figure 5; $P < 1 \times 10^{-2}$), indicating that the DLS was a functional component of each promoter. The *Tesc* promoter contained the initial 'GC' that was found to be most active in previous *in vitro* assays (24). Mutation of the *Tesc* DLS resulted in the greatest loss of function in this analysis, consistent with our previous experimental evidence that a motif with a G at position one bound TFIID most efficiently (24).

Characterization of DLS promoter features

Using the IUPAC consensus motifs from the 'Materials and Methods' section, we mapped all core promoter motifs to all promoters and found that 38% lacked a recognizable motif, [whereas Gershenzon and Ioshikhes (16) reported a slightly lower but comparable amount of 22.7%]. Because DLS could act in combination with other functional core elements to position TFIID, we examined the co-occurrence of other basal promoter elements including TATA, DPE, BRE and CCAAT in the 140 bp region around the TSS. Of the 1621 sequences containing DLS, 1.4% also contained the TATA motif (Table 3). Despite the strict positioning of TATA relative to the TSS site in those promoters, the DLS was positioned throughout the 140 bp region (Supplementary Figure S3). In addition to TATA, 30.2% of DLS promoters contained DPE, 15.6% had Inr, 12.7% had BRE and 8.1% had CCAAT. Overall, 104 promoters contained only a DLS motif with no other core promoter elements, Sp1 motifs or CpG-islands. A GO analysis identified a significant enrichment for the DLS-only promoters regarding the terms calmodulin binding, signal transduction, cell communication and cyclic-nucleotide phosphodiesterase activity (average $P < 1 \times 10^{-2}$) (Supplementary Table S1).

Two additional DLS-like motifs were found in the promoter of *ANK*-1, which is an alternative promoter that specifies erythroid-specific activity in human and mouse cells but lacks sequence conservation in human–mouse genomic alignments (24). Given the non-conserved status of the DLS motif in the *ANK*-1 promoter, our hypothesis was that genomic DLS motifs would not be strongly conserved. However, a number of DLS sites were invariant in primates and showed strong conservation among mammalian clades. Overall this collection of 710 promoters represented genes with enrichment for GO terms of 'anatomical structure development' as found in the brain-specific gene, *POU4F2* and deafness gene, *COL1A2* ($P < 4 \times 10^{-21}$) (Supplementary Table S1). Moreover, 71 DLS motifs were invariant across the five-species alignments (human, chimp, rhesus, mouse and dog), yet did not localize to ultraconserved regions (31) or ultraconserved promoters (i.e. 98% identity for at least 30 bp in seven mammalian species) (32) ('Materials and Methods' section). Genes with invariant DLS sequences showed enrichment for GO terms 'regulation of transcription', 'developmental processes' and 'protein binding' ($P < 1 \times 10^{-5}$) (Supplementary Table S1). DLS genes were compared directly with

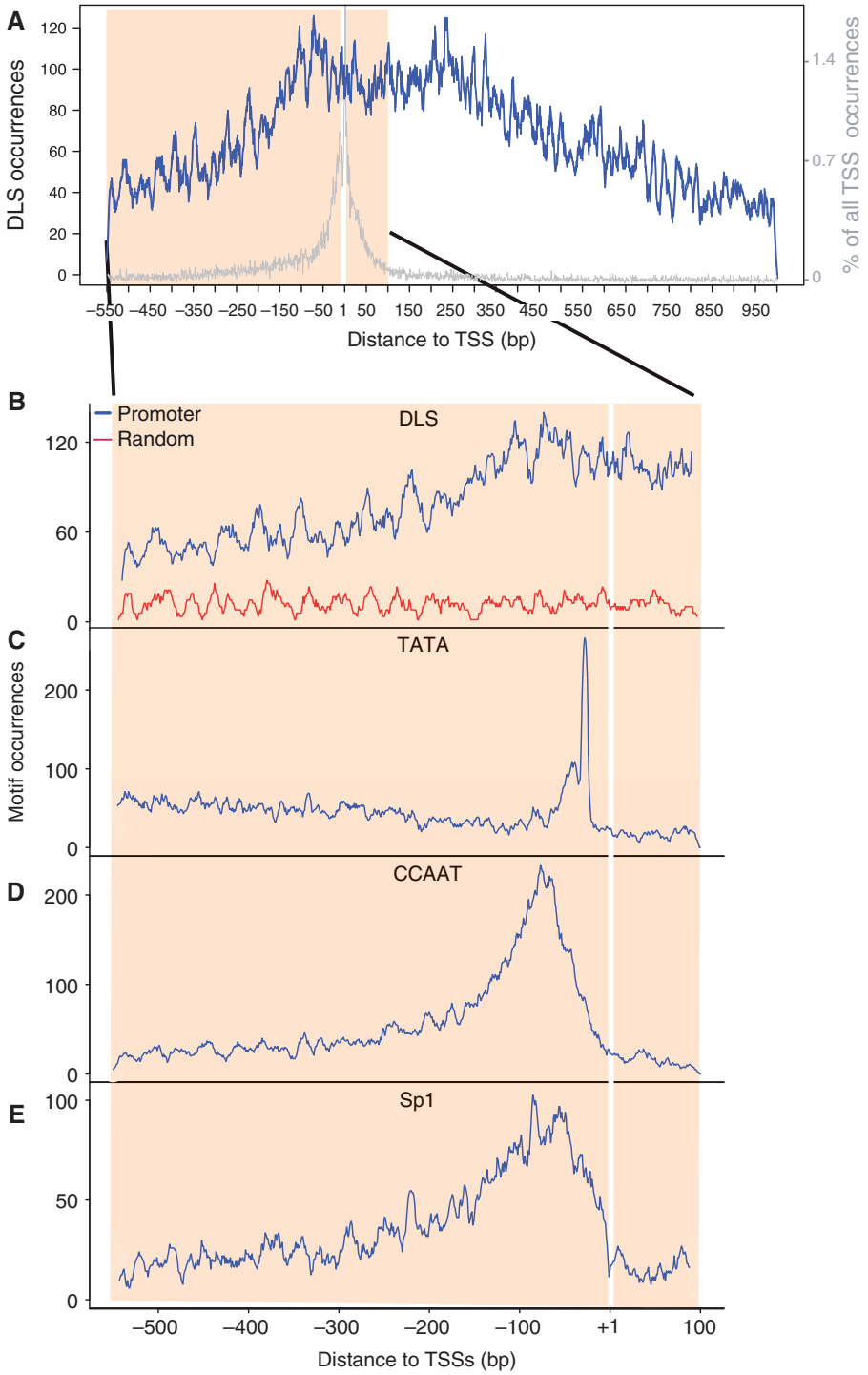


Figure 3. Distribution of the DLS consensus sequence (K)(S)(S)GGNGAG in genomic sequences and control sequences. Every occurrence of a DLS is mapped relative to the TSS identified in the dbTSS database for a region extending from −550 upstream to +1000 downstream (blue line) (**A**). To illustrate the breadth of initiation sites around the +1 position of all reference TSS sites, individual (unclustered) CAGE TSS tags are plotted in gray as a percent of the total number of TSSs present in dbTSS. The orange background indicates windows of the same size shown in (**A–E**). The white line at position +1 represents the position of the primary CAGE TSS annotated for each of the promoter sequences we extracted. Lower panels (B–E) illustrate the region from −550 to +100. DLS occurrences in promoter sequences are plotted in blue and random sequences are plotted in red (B). Positions of TATA, CCAAT and Sp1 motifs are plotted relative to the aggregate TSS position at +1 (C, D, E, respectively).

TATA box genes that demonstrated conservation over the same phylogenetic distances. Significant differences were found for DLS-containing genes, which were over-represented for terms 'zinc ion binding' and 'regulation of transcription', whereas TATA box genes showed enrichment for 'defense response', 'receptor binding' and 'intermediate cytoskeleton filament' ($P < 1 \times 10^{-3}$; Supplementary Table S1). Regarding nucleotide variation

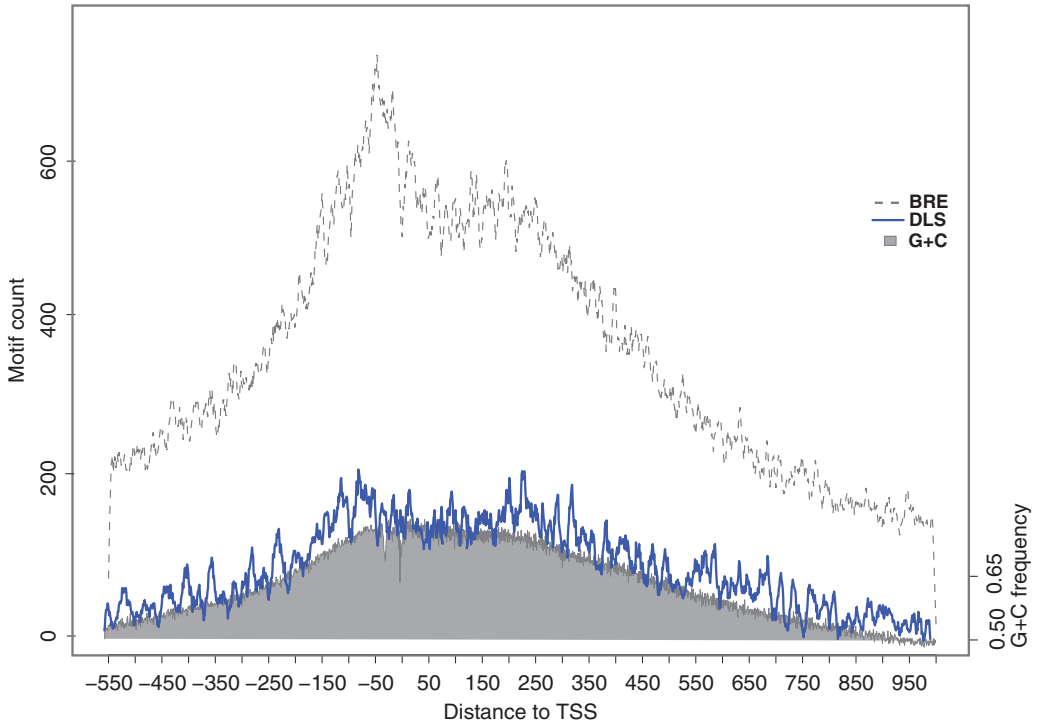


Figure 4. Profiles of DLS, BRE and G+C content in extended promoter sequences. The motif count is plotted relative to the distance to the TSS for BRE (dashed line) and DLS (solid line). The G+C frequency at every position is plotted as the gray area for which the units appear on the right-vertical axis.

Table 2. DLS co-localization with known promoter elements

Promoter subtype	Percentage of 17 181 human promoters	Percentage of DLS promoters	Percentage of promoter subtypes having DLS	Percentage of DLS partially overlapping other motifs
CpG island	60.4	81.5	12.2	–
Non-CpG island	39.6	18.5	4.2	–
DPE	31.3	30.2	8.7	16
Inr	23.1	15.6	8.1	9
BRE	9.3	12.7	12.3	8
CCAAT	8.1	8.1	9	6
TATA	2.8	1.4	4.3	0
DLS	9.0	100	100	–

DLS data were generated using the (K)(S)(S)GGNGAG motif.

Table 3. DLS sites used in mutational analyses

Gene	Motif	Distance upstream of TSS
FBXO18	GGGGCGGAG	53
mKlf3	CCCGGGGAG	11
mTesc	GCGGGCGAG	67
KCC3	CGCGGCGAG	16
Reference motif	(G/T)(C/G)(C/G)GGNGAG	Variable

Underlined sites differ from the consensus.

at Position 6, we found that only 6 of the 71 motifs had a 'T' at this position (Supplementary Figure S5). Human SNPs were found in only one of the 71 invariant mammalian DLS motifs. The promoter of *TMEM180* contained a SNP at DLS Position 7, which was consistently a 'G' in our analysis and became an 'A' with the SNP. Of the genes containing DLS and no other promoter motifs, 20 were involved in a cell death network defined by ingenuity pathways analysis (Supplementary Figure S4).

Although DLS were not strictly positioned relative to the TSS, a potential functional role was addressed through co-occurrence with RNA Pol II binding sites (34), FAIRE (37) and DNaseI hypersensitive sites (HS) in Hela, K562 and GM12878 cells (34). Of 1621 DLS motifs, 71.1% overlapped RNA Pol II ChIP-seq data, 87.9% overlapped DNaseI HS sites and 46.5% coincided with FAIRE sequences. When combined, 89.5% of DLS sequences intersected one or more of these regulatory signals (Figure 6). In contrast, we examined TATA promoters in the human genome and found that 54.6% of the TATA boxes intersected one or more of the RNA Pol II ChIP-seq, DNaseI HS site and FAIRE regions. These results showed that promoters containing DLS were more likely to be active in HeLa, K562 cells and GM12878 cells than were TATA promoters.

Experimental testing of DLS and SP1 as core promoter motifs

Smale *et al.* has proposed that Sp1 directs RNA Pol II to TATA- and InR-like sequences located ~100 bp

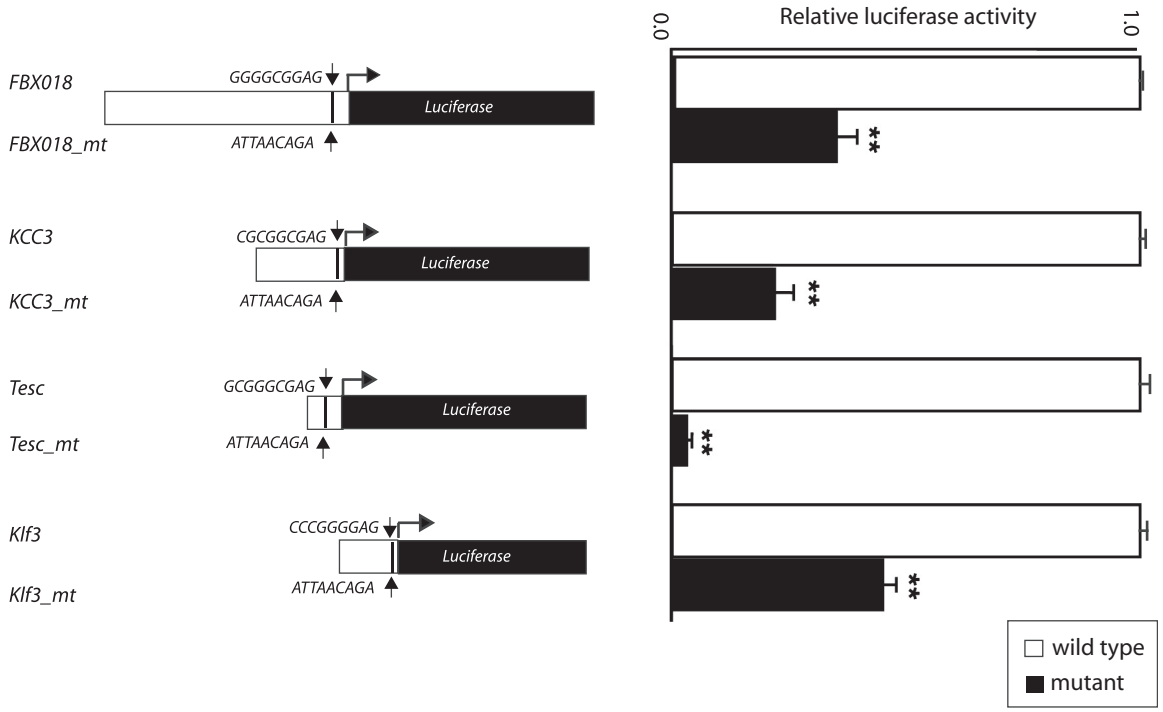


Figure 5. Transient transfection analysis of four promoters containing DLS sequences in K562 cells. Normal DLS sequences were replaced by a random sequence in the expression vectors prior to transient transfection and luciferase assays. The expression vectors are illustrated on the left with sequences of the DLS shown above each image or random sequences shown below each image. The bar plots show the ratio of firefly luciferase expression to *Renilla* luciferase expression for the normal DLS sequences or random sequences, each normalized to the normal DLS ratio; $^{**}P < 0.01$. White bars in the plot represent luciferase expression levels from promoters containing the unaltered DLS sequence, whereas black bars represent the outcome of replacing the DLS motif ('_mt').

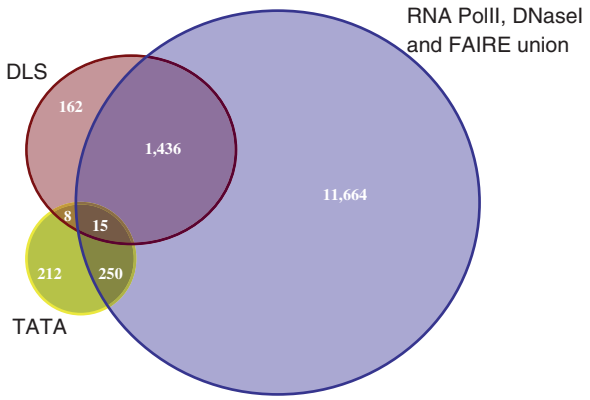


Figure 6. Co-localization of functional features with DLS or TATA. Venn diagram showing the union of DNaseI HS sites, RNA Pol II ChIP-seq and FAIRE data from HeLa K562 and GM12878 cells overlapping by genomic coordinates with DLS or TATA promoters.

downstream (38,39). To determine whether the activity of the DLS was dependent on other core promoter motifs, we constructed a series of plasmids to reconstitute a functional promoter by sequentially adding a DLS motif and Sp1 sites into a promoter-less luciferase reporter-vector. In this assay, three Sp1 binding sites $(Sp1)_3$ alone increased luciferase activity 9-fold compared with the promoter-less plasmid, indicating that Sp1 sites conferred basal promoter activity (Figure 7). In contrast, a plasmid containing one copy of the DLS motif was not active. However, the addition of a DLS motif 100 bp downstream of the Sp1 sites, $[(Sp1)_3$+DLS], increased activity 2-fold over the $(Sp1)_3$ ($P < 0.01$). Mutation of the DLS motif reduced activity to the level of $(Sp1)_3$ alone ($P < 0.01$), confirming a functional contribution of the DLS.

DISCUSSION

The identification of new promoter elements in the human genome is an ongoing process. TATA boxes were initially defined by their tight adherence to a consensus sequence, a biased position relative to the TSS and AT-richness of promoters in prokaryotic and early eukaryotic genomes (40). A variety of core promoter elements have been identified in vertebrates, however a significant number of human promoters lack any recognizable promoter elements. Completion of the human genome-sequencing project has revealed that consensus TATA box motifs comprise only a minority of all human promoter sequences. Other basal elements occur with both larger and smaller frequencies than TATA indicating that novel elements may occur in a very small, but meaningful proportion of human promoters. These elements may not be fully elucidated until all promoters are comprehensively identified in the genome.

Based on biological consequences of a disease mutation, we predicted that the TFIID localization signal present in a 9 bp sequence of the ANK-1 promoter represented a

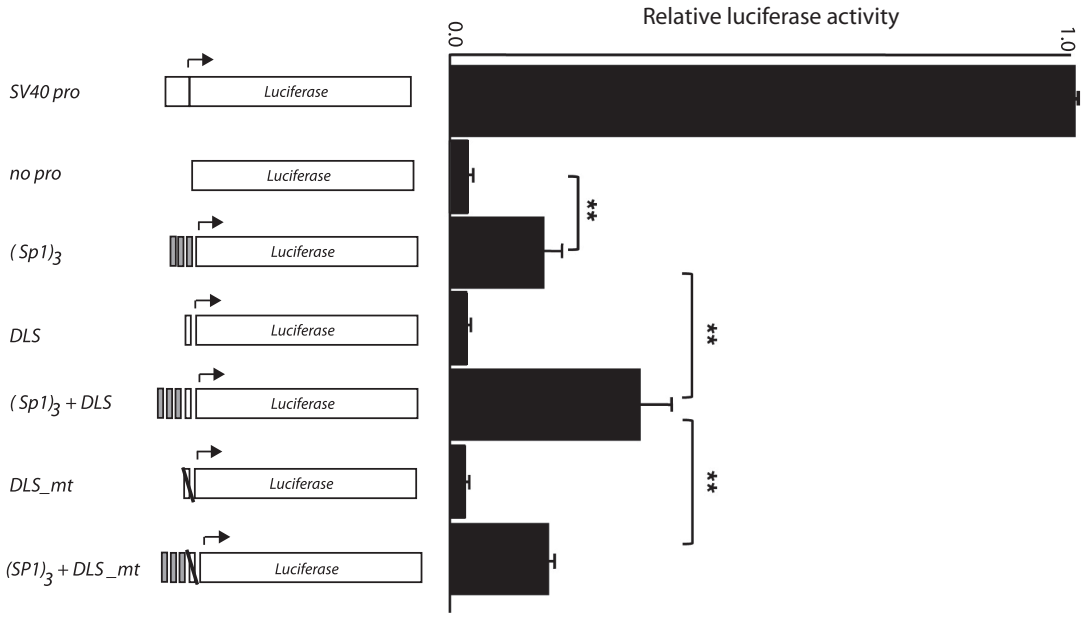


Figure 7. Transient transfection analysis of synthetic promoters in K562 cells. The data are from the ratio of firefly luciferase expression/*Renilla* expression for each plasmid, normalized to the ratio generated by the plasmid carrying the SV40 promoter only. Each plasmid contains a specific promoter motif with or without a mutation or no motif at all: promoter-less (no pro), three Sp1 binding sites $(Sp1)_3$, consensus DLS sequence (DLS), mutant DLS sequence (DLS_mt). Statistical significance was assessed using a student's *t*-test ($^{**}P < 0.01$).

functional TFIID localization element for promoters throughout the genome and using mutational analyses, confirmed this functional role. We found a significant enrichment of the DLS motif in human promoter sequences compared with background models, with a greater representation in CpG-islands. The enriched presence of DLS in CpG-island promoters was not a chance event due to high G+C nucleotide content in the promoter sequence, as shown by significant differences in the DLS occurrence in promoter sequences and background models created by 1 and 2 bp sampling. Furthermore the DLS was enriched in CpG-islands of core promoters compared with CpG-islands found elsewhere in the genome, indicating that the enrichment of DLS motif was specific to core promoter regions rather than CpG-islands. No preferential positioning of the DLS was observed relative to the TSSs of the associated transcripts, regardless of whether the promoter contained a TATA motif or was devoid of any other core elements. Despite the fact that closely spaced TSSs have been clustered into single representative TSSs in the CAGE data (Firth *et al*. 2008), the profile of DLS motifs was wider than the profile of all initiation sites (mapped as unclustered CAGE tags) overlapping those promoters. Therefore the broad-peak initiation status of the CpG-island promoters did not provide an explanation for the breadth of the DLS distribution. We cannot rule out the hypothesis that the DLS may play a role in positioning TFIID from downstream locations [as do MTE and DPE (41)], while also functioning from upstream positions, since motif-specific targeting mechanisms at broad peak promoters have not been elucidated (17).

A model has been proposed in which TFIID is recruited to CpG-rich core promoter regions through Sp1 directly contacting the DNA. This TFIID PIC functionally resembles those formed at either TATA or Inr sequences without Sp1 (33,42). Our data demonstrate that Sp1 is capable of supporting basal promoter activity and like Sp1 the DLS element is an additional sequence that can be nested within CpG-islands to contribute further regulatory function. Combined with our previous data showing that the presence of a deletion in this motif causes a human disease (24), we present evidence that the DLS motif is found frequently in human promoters, loses function when mutated *in vitro* (and in an *in vivo* deletion mutation) and increases promoter activity when combined with Sp1 sites. These properties are consistent with the role of a core promoter element in the human genome (16). Nevertheless, the DLS is inconsistent with a core promoter motif because it does not have a positional preference relative to the TSS and does not act alone. We conclude the DLS motif contains a functional promoter element that has a significant role in human gene regulation. Our approach to identifying this motif was based on an observation of a human mutation. A compendium of human mutations that reveal functional sites within core promoters, such as single nucleotide substitutions in thalassemias, confirms that analyses of these types of mutations have been useful for defining novel instances of promoter motifs (43).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Bruno Salvy for help with calculations of theoretical expectations based on nucleotide frequencies in core promoters.

FUNDING

NHGRI intramural funds (to L.E. and D.M.B.), DK60239, DK04015 (to P.G.G.); Fonds de la recherche en santé du Québec (to K.L.); Ohio University Stocker Endowment, Ohio University's Graduate Research and Education Board (GERB); Ohio Supercomputer Center (to L.W., J.L.). Funding for open access charge: NHGRI intramural funds.

Conflict of interest statement. None declared.

REFERENCES

1. Smale,S.T. (2001) Core promoters: active contributors to combinatorial gene regulation. *Genes Dev.*, **15**, 2503–2508.
2. Gross,P. and Oelgeschlager,T. (2006) Core promoter-selective RNA polymerase II transcription. *Biochem. Soc. Symp.*, 225–236.
3. Smale,S.T. and Kadonaga,J.T. (2003) The RNA polymerase II core promoter. *Annu. Rev. Biochem.*, **72**, 449–479.
4. Butler,J.E. and Kadonaga,J.T. (2002) The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev.*, **16**, 2583–2592.
5. Zhang,M.Q. (1998) A discrimination study of human core-promoters. *Pac. Symp. Biocomput.*, 240–251.
6. Lewis,B.A., Kim,T.K. and Orkin,S.H. (2000) A downstream element in the human beta-globin promoter: evidence of extended sequence-specific transcription factor IID contacts. *Proc. Natl Acad. Sci. USA*, **97**, 7172–7177.
7. Antequera,F. and Bird,A. (1994) Predicting the total number of human genes. *Nat. Genet.*, **8**, 114.
8. Suzuki,Y., Tsunoda,T., Sese,J., Taira,H., Mizushima-Sugano,J., Hata,H., Ota,T., Isogai,T., Tanaka,T., Nakamura,Y. *et al.* (2001) Identification and characterization of the potential promoter regions of 1031 kinds of human genes. *Genome Res.*, **11**, 677–684.
9. Juven-Gershon,T., Hsu,J.Y. and Kadonaga,J.T. (2006) Perspectives on the RNA polymerase II core promoter. *Biochem. Soc. Trans.*, **34**, 1047–1050.
10. O'Shea-Greenfield,A. and Smale,S.T. (1992) Roles of TATA and initiator elements in determining the start site location and direction of RNA polymerase II transcription. *J. Biol. Chem.*, **267**, 6450.
11. Emami,K.H., Jain,A. and Smale,S.T. (1997) Mechanism of synergy between TATA and initiator: synergistic binding of TFIID following a putative TFIIA-induced isomerization. *Genes Dev.*, **11**, 3007–3019.
12. Burke,T.W. and Kadonaga,J.T. (1997) The downstream core promoter element, DPE, is conserved from Drosophila to humans and is recognized by TAFII60 of Drosophila. *Genes Dev.*, **11**, 3020–3031.
13. Zhou,T. and Chiang,C.M. (2001) The intronless and TATA-less human TAF(II)55 gene contains a functional initiator and a downstream promoter element. *J. Biol. Chem.*, **276**, 25503–25511.
14. Tsai,F.T. and Sigler,P.B. (2000) Structural basis of preinitiation complex assembly on human pol II promoters. *EMBO J.*, **19**, 25–36.
15. Lagrange,T., Kapanidis,A.N., Tang,H., Reinberg,D. and Ebright,R.H. (1998) New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. *Genes Dev.*, **12**, 34–44.
16. Gershenzon,N.I. and Ioshikhes,I.P. (2005) Synergy of human Pol II core promoter elements revealed by statistical sequence analysis. *Bioinformatics*, **21**, 1295–1300.
17. Muller,F., Demeny,M.A. and Tora,L. (2007) New problems in RNA polymerase II transcription initiation: matching the diversity of core promoters with a variety of promoter recognition factors. *J. Biol. Chem.*, **282**, 14685–14689.
18. Juven-Gershon,T., Hsu,J.Y., Theisen,J.W. and Kadonaga,J.T. (2008) The RNA polymerase II core promoter - the gateway to transcription. *Curr. Opin. Cell Biol.*, **20**, 253–259.
19. Frith,M.C., Valen,E., Krogh,A., Hayashizaki,Y., Carninci,P. and Sandelin,A. (2008) A code for transcription initiation in mammalian genomes. *Genome Res.*, **18**, 1–12.
20. Sandelin,A., Carninci,P., Lenhard,B., Ponjavic,J., Hayashizaki,Y. and Hume,D.A. (2007) Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat. Rev. Genet.*, **8**, 424–436.
21. Brandeis,M., Frank,D., Keshet,I., Siegfried,Z., Mendelsohn,M., Nemes,A., Temper,V., Razin,A. and Cedar,H. (1994) Sp1 elements protect a CpG island from de novo methylation. *Nature*, **371**, 435–438.
22. Blake,M.C., Jambou,R.C., Swick,A.G., Kahn,J.W. and Azizkhan,J.C. (1990) Transcriptional initiation is controlled by upstream GC-box interactions in a TATAA-less promoter. *Mol. Cell. Biol.*, **10**, 6632–6641.
23. Li,L., He,S., Sun,J.M. and Davie,J.R. (2004) Gene regulation by Sp1 and Sp3. *Biochem. Cell Biol.*, **82**, 460–471.
24. Laflamme,K., Owen,A.N., Devlin,E.E., Yang,M.Q., Wong,C., Steiner,L.A., Garrett,L.J., Elnitski,L., Gallagher,P.G. and Bodine,D.M. (2010) Functional analysis of a novel cis-acting regulatory region within the human ankyrin (ANK-1) gene promoter. *Mol. Cell. Biol.*, **30**, 3493–3502.
25. Gallagher,P.G., Romana,M., Tse,W.T., Lux,S.E. and Forget,B.G. (2000) The human ankyrin-1 gene is selectively transcribed in erythroid cell lines despite the presence of a housekeeping-like promoter. *Blood*, **96**, 1136–1143.
26. Gallagher,P.G., Sabatino,D.E., Basseres,D.S., Nilson,D.M., Wong,C., Cline,A.P., Garrett,L.J. and Bodine,D.M. (2001) Erythrocyte ankyrin promoter mutations associated with recessive hereditary spherocytosis cause significant abnormalities in ankyrin expression. *J. Biol. Chem.*, **276**, 41683–41689.
27. Gallagher,P.G., Nilson,D.G., Wong,C., Weisbein,J.L., Garrett-Beal,L.J., Eber,S.W. and Bodine,D.M. (2005) A dinucleotide deletion in the ankyrin promoter alters gene expression, transcription initiation and TFIID complex formation in hereditary spherocytosis. *Hum. Mol. Genet.*, **14**, 2501–2509.
28. Suzuki,Y., Yamashita,R., Nakai,K. and Sugano,S. (2002) DBTSS: DataBase of human transcriptional start sites and full-length cDNAs. *Nucleic Acids Res.*, **30**, 328–331.
29. Nicodème,P., Salvy,B. and Flajolet,P. (2002) Motif statistics. *Theoret. Comput. Sci.*, **287**, 593–618.
30. Schbath,S., Prum,B. and Tur-ckheim,E.d. (1995) Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences. *J. Comput. Biol.*, **2**, 417–437.
31. Bejerano,G., Pheasant,M., Makunin,I., Stephen,S., Kent,W.J., Mattick,J.S. and Haussler,D. (2004) Ultraconserved elements in the human genome. *Science*, **304**, 1321–1325.
32. Rodelsperger,C., Kohler,S., Schulz,M.H., Manke,T., Bauer,S. and Robinson,P.N. (2009) Short ultraconserved promoter regions delineate a class of preferentially expressed alternatively spliced transcripts. *Genomics*, **94**, 308–316.
33. Deng,W. and Roberts,S.G. (2006) Core promoter elements recognized by transcription factor IIB. *Biochem. Soc. Trans.*, **34**, 1051–1053.
34. Boyle,A.P., Davis,S., Shulha,H.P., Meltzer,P., Margulies,E.H., Weng,Z., Furey,T.S. and Crawford,G.E. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311–322.
35. Antequera,F. (2003) Structure, function and evolution of CpG island promoters. *Cell. Mol. Life Sci.*, **60**, 1647–1658.
36. 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.
37. Giresi,P.G., Kim,J., McDaniell,R.M., Iyer,V.R. and Lieb,J.D. (2007) FAIRE (formaldehyde-assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin. *Genome Res.*, **17**, 877–885.
38. Smale,S.T., Schmidt,M.C., Berk,A.J. and Baltimore,D. (1990) Transcriptional activation by Sp1 as directed through TATA or initiator: specific requirement for mammalian transcription factor IID. *Proc. Natl Acad. Sci. USA*, **87**, 4509–4513.

39. Carey,M. and Smale,S.T. (2000) *Transcriptional regulation in euckaryotes, concepts, strategies and techniques*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
40. Khuu,P., Sandor,M., DeYoung,J. and Ho,P.S. (2007) Phylogenomic analysis of the emergence of GC-rich transcription elements. *Proc. Natl Acad. Sci. USA*, **104**, 16528–16533.
41. Theisen,J.W., Lim,C.Y. and Kadonaga,J.T. (2010) Three Key Subregions Contribute to the Function of the Downstream RNA Polymerase II Core Promoter. *Mol. Cell. Biol.*, **30**, 3471–3479.
42. Smale,S.T. (1994) Core promoter architecture for eukaryotic protein-coding genes. In Conaway,R.C. and Conaway,J.W. (eds), *Transcription: Mechanisms and regulation*. Raven Press, Ltd., New York, pp. 63–81.
43. Savinkova,L.K., Ponomarenko,M.P., Ponomarenko,P.M., Drachkova,I.A., Lysova,M.V., Arshinova,T.V. and Kolchanov,N.A. (2009) TATA box polymorphisms in human gene promoters and associated hereditary pathologies. *Biochemistry*, **74**, 117–129.