A Genome Scale Location Analysis of Human Staf/ZNF143-binding Sites Suggests a Widespread Role for Human Staf/ZNF143 in Mammalian Promoters

Evelyne Myslinski, Marie-Aline Gérard, Alain Krol, and Philippe Carbon

From the Institut de Biologie Moléculaire et Cellulaire, UPR CNRS Architecture et Réactivité de l’ARN, Université Louis Pasteur, 15 Rue René Descartes, 67084 Strasbourg Cedex, France

Staf was originally identified as the transcriptional activator of Xenopus tRNA\textsuperscript{Sec} and small nuclear (sn) RNA-type genes. Recently, transcription of seven human (h) protein coding genes was reported to be activated by the human ortholog hStaf/ZNF143. Here we have used a combined \textit{in silico} and biochemical approach to identify 1175 conserved hStaf/ZNF143-binding sites (SBS) distributed in 938 promoters of four mammalian genomes. The SBS shows a significant positional preference and occurs mostly within 200 bp upstream of the transcription start site. Chromatin immunoprecipitation assays with 295 of the promoters established that 90% contain \textit{bona fide} SBS. By extrapolating the values of this mapping to the full sizes of the mammalian genomes, we can infer the existence of at least 2500 SBS distributed in 2000 promoters. This unexpected large number strongly suggests that SBS constitutes one of the most widespread transcription factor-binding sites in mammalian promoters. Furthermore, we demonstrated that the presence of the SBS alone is sufficient to direct expression of a luciferase reporter gene, suggesting that hStaf/ZNF143 can recruit \textit{per se} the transcription machinery.

The vertebrate transcription factor Staf was originally identified in \textit{Xenopus laevis} as the transcriptional activator of the tRNA\textsuperscript{Sec} gene (1). Since then, Staf has also been involved in transcriptional activation of snRNA\textsuperscript{2} and snRNA-type genes (2, 3). Staf binds to the Staf-binding site (SBS), a sequence identified in promoters of many vertebrate snRNA and snRNA-type genes transcribed by RNA polymerases II or III. Seven contiguous zinc fingers of the C2-H2 type, located in the central part of the protein, contain the DNA-binding domain. Binding site selection experiments identified the 18-bp YYWCCCCRN-MATSCMYYRCR (Y, W, R, N, M, and S stand for T/C, A/T, A/G, any nucleotide, A/C, and G/C, respectively) or TACCCATAATGATYGGCG sequences as the Staf consensus binding sites, depending on whether moderate or highly stringent conditions were employed (2, 4). Compared with the consensus sequences, known Staf-binding sites revealed a high degree of divergence. This is well illustrated by the absence of either the 5’ part in the \textit{Xenopus} tRNA\textsuperscript{Sec} or the 3’ part in the human U6 snRNA sites (5). Not all the seven zinc fingers are necessary for the binding of Staf to the SBS. For example, zinc finger 7 does not establish base-specific contacts in Staf-DNA complexes, and the requirement for zinc finger 1 is flexible because it contacts the DNA at the \textit{Xenopus} tRNA\textsuperscript{Sec} but not human U6 snRNA Staf motifs (5). The nonutilization of zinc finger 1 at the human U6 promoter enables the simultaneous binding of Staf and Oct-1 to their cognate DNA motifs; Oct-1 is another factor involved in transcriptional activation of the human U6 snRNA gene (5).

ZNF76 and ZNF143 are two human homologs of \textit{Xenopus} Staf, with ZNF143 being the ortholog and ZNF76 a likely paralog (6–8). ZNF143 is conserved in vertebrates and urochordates, such as \textit{Ciona intestinalis}, whereas ZNF76 is only present in vertebrates. In this study, human Staf will be referred to as hStaf/ZNF143.

Moreover, Staf possesses the capacity to stimulate chloramphenicol acetyltransferase expression from a synthetic mRNA promoter (SBS-tkCAT), containing SBS linked to the basal promoter of the thymidine kinase gene fused to the chloramphenicol acetyltransferase reporter (9). The presence of two physically and functionally distinct activation domains constitutes the molecular basis underlying the abilities of hStaf/ZNF143 to stimulate transcription from either snRNA-type or mRNA promoters. Indeed, although a 93-amino acid domain made of four repeat units is specialized for activating transcription from an mRNA promoter, a segment of only 18 amino acids acts specifically on snRNA and snRNA-type promoters (10). To date, only the following seven protein coding genes have been described to be regulated by hStaf/ZNF143: the cytosolic chaperonin containing t-complex polypeptide 1 (TCP1) (11); the interferon regulatory factor (IRF3) (12); the neuronal nitric-oxide synthase (NOS1) (13); the transaldolase (TALDO1) (14); the aldheyde reductase (AKR1A1) (15); the mitochondrial ribosomal protein S11 (MRPS11) (16); and the synaptobrevin-like 1 (SYBL1) (17).
**hStaf/ZNF143-binding Sites in the Human Genome**

In this work, we describe the results of an analysis aimed at identifying hStaf/ZNF143-binding sites across mammalian genomes. We have used *in silico* genome-wide analysis to identify 1175 SBS distributed in 938 promoters, among which 295 were chosen for experimental validation by chromatin immuno-precipitation (ChIP). Semi-quantitative PCR confirmed the binding of hStaf/ZNF143 to 266 (90%) of the tested loci. Most of the identified SBS were located near the transcription start site and were tightly associated with CpG islands. Extrapolated to the full sizes of the mammalian genomes, this partial mapping revealed an unexpected large number of SBS in the promoters of protein coding genes, with at least 2500 SBS distributed in 2000 promoters. Furthermore, this study established that the presence of SBS alone is sufficient to stimulate the expression of a luciferase reporter gene, suggesting that hStaf/ZNF143 can recruit the transcription machinery and up-regulate transcription of an extraordinary high number of genes.

**EXPERIMENTAL PROCEDURES**

**Bioinformatics**—Human promoters containing putative SBS were obtained by searching transcription factor-binding sites at DBTSS (18, 19). DBTSS version 4 (2003 human hg16 assembly) was queried with the Staf-binding sites listed in Table 1. Alignments of the human and mouse promoter sequences provided by DBTSS were examined for Staf-binding site conservation in the mouse. Using the human promoter sequences containing the SBS motif as the query, we looked for motif conservation in mouse (for promoters not present in DBTSS), rat, and dog genomes based on the vertebrate Multiz alignment (BLAT search at the UCSC Genome Bioinformatics Site). Assignment of the identified genes to categories of functional groups was performed using information from NCBI Entrez gene and Gene Ontology provided by GOA.

The proximity of CpG islands to the identified SBS was found by examining at DBTSS the results of the search for transcription factor-binding sites. The coordinates of the human Staf-binding sites (based on the May 2004 human hg17) were extracted by BLAT search at the UCSC Genome Bioinformatics Site, using the corresponding sequence. Positions of the Staf-binding sites in human and mouse promoters were obtained by searching DBTSS for transcription factor-binding site.

Identification of additional Staf-binding sites from the work of Xie et al. (20) was performed as follows. The coordinates of the discovered motifs corresponding to sequences ACTAY-RNNCCCR, ACTACNNNCCCC, and ACTACNNNTCCCR (Xie motifs), in direct and reverse orientations, were obtained on line. Sequences located 3’ to the Xie motifs were extracted with the Human Genome Browser. 3’-Extended Xie motifs were visually examined for the presence of an associated Staf-binding site. Conservation of the newly identified Staf motifs, in the mouse, rat, and dog genomes, was evaluated by the vertebrate Multiz alignment. Sequence logos were created using WebLogo (21).

**Reporter Constructs**—WTSBS-Luc, mut-1 SBS-Luc, and mut-2 SBS-Luc were obtained by inserting into the promoterless pGL3-basic vector (Promega) a 150-bp KpnI-HindIII DNA fragment containing three copies of the tRNA*	ext{Sec}* gene SBS, either the WT GTACCAGCATGCGCTCAGCG, mut-1 GTACC-ATACGTACTCGCG, or mut-2 GTAAACGCATGCCT-CGCG sequences, respectively (1, 9).

**Transfection and Luciferase Assays**—CO-S7 cells were transfected by the calcium phosphate co-precipitation procedure with 1 µg of reporter construct, 0.5 µg of pCH110 plasmid as the internal control, and carrier DNA to bring up the total DNA content to 10 µg/plate. After 24 h, cells were lysed, and the β-galactosidase activity was measured as described previously (10). The luciferase assay was performed as recommended by the manufacturer. The luciferase activity was normalized to the β-galactosidase activity. Each transfection experiment was done in triplicate.

**ChIP Assay**—The ChIP procedure was essentially as described in Weinmann and Farnham (22) with a few modifications. HeLa cells (10⁶ cells/ml) were treated with 1% formaldehyde at room temperature for 10 min. The reaction was stopped by addition of glycine to a final concentration of 125 mM. Cells were washed once with ice-cold phosphate-buffered saline containing protease inhibitors, scraped, centrifuged at 1500 rpm for 10 min, then resuspended in cell lysis buffer (5 mM HEPES-NaOH, pH 8, 85 mM KCl, 0.5% Nonidet P-40, protease inhibitors), and kept on ice for 10 min. They were homogenized 10 times with a Dounce homogenizer, and the resultant homogenates were centrifuged at 4000 rpm for 10 min to pellet the nuclei. Nuclei were resuspended in 1.5 ml of shearing buffer (50 mM Tris-HCl, pH 8, 10 mM EDTA, 1% SDS, protease inhibitors). Nuclear lysates were sonicated on ice to an average chromatin length of 0.5–1 kb (three pulses of 20 s at a setting of 25 with a Vibra-cell sonicator; Bioblock Scientific) and then centrifuged at 14,000 rpm for 12 min. The supernatant was incubated with 10 volumes of ChIP buffer (50 mM HEPES-NaOH, pH 7.9, 140 mM NaCl, 1% Triton X-100, 0.01% SDS, protease inhibitors) with protein A-Sepharose for 2 h at 4 °C. After protein A-Sepharose removal, the precleared lysates were used as the soluble chromatin for ChIP. Chromatin was incubated at 4 °C overnight with 3.5 µg of anti-hStaf/ZNF143. Control for no antibody was included. Immune complexes were recovered by incubation for 2 h at 4 °C with protein A-Sepharose previously blocked in buffer containing 2 mg/ml bovine serum albumin and 1 mg/ml salmon testis DNA. Immune complexes were washed four times sequentially in the following buffers (shearing buffer; wash buffer 1: 50 mM HEPES-NaOH, pH 7.9, 300 mM NaCl, 1% Triton X-100, 0.1% SDS; wash buffer 2: 50 mM HEPES-NaOH, pH 7.9, 500 mM NaCl, 1% Triton X-100, 0.1% SDS; wash buffer 3: 50 mM Tris-HCl, pH 8, 250 mM LiCl, 0.5% Nonidet P-40, 1 mM EDTA, 1% deoxycholate; TE buffer). The DNA-protein complex was eluted in 400 µl of elution buffer (100 mM NaHCO₃, 1% SDS) at room temperature for 15 min. The formaldehyde cross-link was reversed by overnight incubation at 65 °C in the presence of 190 mM NaCl. Immunoprecipitated DNA was released by treatment with 100 µg of proteinase K for 1 h at 45 °C in the presence of 10 mM EDTA and 40 mM Tris-HCl, pH 6.5. The DNA was purified by phenol extraction and ethanol-precipitated. The rabbit polyclonal antipeptide antibody against a C-terminal epitope of the *Xenopus* Staf protein (5) was used for ChIP.

**PCR Analysis**—Purified DNA was analyzed in 25-µl PCRs with primer pairs flanking the predicted SBS and in the pres-
RESULTS

In Silico Identification of 893 Potential hStaf/ZNF143-binding Sites across Human and Mouse Protein Gene Promoters—Our strategy for identifying human promoters directly bound by Staf was based on the pre-selection of candidates showing phylogenetic conservation of the SBS site in orthologous mammalian promoters, followed by the experimental analysis of a large number of individual sites with a semi-quantitative ChIP assay. The pre-selected promoters were identified by searching potential human transcription factor-binding sites at the DBTSS promoter data base (18, 19) that are conserved in orthologous mouse promoters. Version 4 used in this study was originally constructed from a collection of experimentally determined transcription start sites (TSS) of human genes and is suitable for comparative analysis of human and mouse orthologous promoters. Version 4 used in this study was based on human hg16 and contains 8793 human promoter sequences. From the scrutiny of seven known target genes, we could localize the Staf-binding sites in promoters lying at the proximity of TSS. Therefore, the analysis was further limited to sequences spanning 1000 bp upstream and 200 bp downstream of the TSS. Eleven 18-bp-long sequences (A1 to K1 in Table 1) were used as queries for the first round of motif search. The 18-bp YYCCANNRNCNNYCR A1 sequence is the consensus derived by comparing the eight Staf motifs identified in seven human and mouse protein coding genes (11–17). It is more degenerated than the Staf consensus sequences derived from binding site selection experiments under moderately stringent conditions (see Introduction). With respect to A1, the B1 to K1 sequences bear 1-, 2-, or 3-bp variations (Table 1). These changes were chosen from the constraints observed at several positions in binding site selection experiments (2, 4).

Among the DBTSS entries, a high number of promoters scored positive with one or more potential Staf-binding sites within the predetermined boundaries, identifying a total of 1488 potential binding sites with 648 motifs (43.5%) conserved in the orthologous mouse promoters (Table 1 and supplemental Table S1). Table 1 indicates that the score distribution is not uniform, the highest ones being obtained for the A1 and F1 sequences with 185 and 97 human/mouse conserved hits, the lowest score arising from the H1 sequence with 13 human/mouse conserved hits. Generally, a good correlation was observed at one given position between score distribution and the nucleotide preference obtained from binding site selection experiments (2, 4). For example, binding site selection indicated that a C occupied position 3 in 95% of the sequences, whereas a T occurred in only 3% of the cases (2). In the DBTSS screen, the same distribution was observed. The query sequence carrying a C at position 3 (A1) produced 185 conserved hits between human and mouse, a single C to T change (H1) leading to only 13 hits (Table 1). Interestingly, sequence alignments of putative SBS show that residues preferentially occupying positions 7–11, 13–16, and 18 (Fig. 1A, right panel), corresponding to N, R, or Y in the queried sequences in Table 1, are identical to those derived from the binding site selection experiments (2). This strongly suggests that the identified sequences represent bona fide SBS.

Inspection of the adjacent sequences to the potential SBS revealed the presence of the 7-bp ACTACA motif, or a 1–2–bp variant, among 305 (47%) of the 648 identified SBS (Table 1; Fig. 1A, left panel; supplemental Table S1). It lies immediately 5’ to the SBS. Therefore, to identify novel SBS that would harbor the ACTACA motif, we performed a second round of motif search at DBTSS with the A2 to F2, H2, and J2 motifs as queries. These 19-bp motifs are arranged in the following way (see Table 2). The 5’ part is now composed of the 7-bp submotif ACTACA, the 3’ part corresponding to bp 1–12 of the A1 to F1, H1, and J1 sequences. This second round generated 459 hits with 183 occurrences conserved in the orthologous mouse promoters but not identified by the A1 to K1 queries. It is striking to observe that the consensus sequence derived for the 6-bp sequence directly adjacent to the 3’ part of the newly identified motifs, although not included in the queries, bears high sequence similarity to sequence CC(C/T)GCG of the A1 to K1 motifs (compare positions 13–18 in Fig. 1, A and B; note that the heights of the letters are not directly proportional to occurrences). Again, this strongly argues for the characterization of genuine binding sites. Next, to characterize promoters containing at least two potential SBS, we performed a third search at DBTSS (LA1–LK1, Table 3). This round was implemented with two simultaneous queries: (i) one with the previous A1-K1 motifs (sequence motif 1 in Table 3); (ii) the

---

hStaf/ZNF143-binding Sites in the Human Genome

---

Table 1

Identification of Staf-binding sites by searching transcription factor-binding sites at DBTSS

Query names, sequences of the query, and hits of human versus mouse SBS conservation are indicated. In addition, this table displays the percentages of the ACTACA and variant motifs linked to the 5’ part of the identified SBS. N, R, and Y indicate any nucleotide, A or G, and T or C, respectively.

| Query name | Sequence | Human/mouse hits | ACTACA and variants |
|------------|----------|------------------|---------------------|
| A1         | YYCCANRNCNNYCR | 185              | 61                  |
| B1         | R--        | 31               | 29                  |
| C1         | R--        | 79               | 45                  |
| D1         | RR--------- | 93               | 38                  |
| E1         | R------------ | 17               | 0.5                 |
| F1         | --Y-------- | 97               | 34                  |
| G1         | --Y-------- | 32               | 53                  |
| H1         | --T-------- | 13               | 76                  |
| I1         | --Y-------- | 47               | 63                  |
| J1         | --Y-------- | 16               | 12                  |
| K1         | R------------ | 38               | 38                  |
| A1–K1      | 648       | 47               |                      |
promoters were visually examined for unidentified sequence elements that might serve as Staf-binding sites. This inspection enabled detection of 41 additional sites conserved in orthologous mouse promoters (named X in Fig. 1D and supplemental Table S1). To extend the phylogenetic comparison, a comparative analysis was performed in the orthologous promoters of the rat and dog genomes. Among the 893 identified motifs that we found conserved in the human and mouse genomes, 772 orthologous regions were detected in the rat and canine genomes. Interestingly, sequence comparisons revealed that 753 (97%) are also conserved in these two genomes. In 18 cases, the SBS is conserved in one of the two genomes, and in one single case only (NM_003085) they both lack it (supplemental Table S1).

Altogether, our comparative search identified 893 SBS conserved in mouse and residing in 716 promoters of human protein coding genes. Furthermore, nearly about one-half of the discovered SBS were associated with the newly identified ACTACA\textsubscript{N} motif. Both the SBS and the associated ACTACA\textsubscript{N} motifs are also conserved with a very high prevalence in the rat and dog genomes.

**Identification of 282 Other Putative hStaf/ZNF143-binding Sites**

A recent report described the systematic discovery of regulatory motifs in human promoters by comparing several mammalian genomes (20). Among the 174 conserved motifs that the authors discovered in the promoters of protein coding genes, the 13-bp ACTA\textsubscript{YRN} sequence is particularly interesting because ACTA\textsubscript{YRN} shows a strong match to the SBS-adjacent ACTACA\textsubscript{N} submotif, and \textsubscript{NNCC}R is highly similar to the 5'/H11032 part of the SBS. We thus sought to use the M4 motif to uncover new SBS by searching for the presence of SBS overlapping the M4 motif. This led us to discover 282 new potential SBS (see logo sequence derived from sequence comparisons, Fig. 1E).

Among these, 23 are located in promoters previously identified by our search at DBTSS, and the other 259 SBS are located in 222 new promoters containing this site (Fig. 2A). The 282 sites

![Motif analysis of SBS with sequence logos depicting the nucleotide distribution of the SBS (right panel) and the newly discovered 5'-associated submotifs (left panel). A–C, sequence logo based on the results of the A1–K1, A2–F2, H2, J2, and LA1-LK1 queries. D, sequence logo based on the X sequences. E, sequence logo based on the SBS that were identified linked to the M4 motif of Xie et al. (20).](image-url)
Collectively, our in silico analysis resulted in the identification of 1175 conserved putative SBS harbored in 938 promoters of mammalian genomes. The very high degree of sequence conservation of the six potential SBS appears clearly and stands out from the nonconserved flanking sequences.

Two interesting features regarding the physical arrangement of the SBS came from our study. First, of the 1175 identified SBS, a bias occurs toward a direct orientation with respect to the gene as follows: 702 (60%) are in direct orientation; among promoters possessing two SBS, 41% are in direct orientation, and 13% are in reverse orientation, and the remainder have a combination of both. Second, we identified SBS in potential bidirectional promoters, with 146 promoters (15%) arranged in a head-to-head configuration within 1 kbp (supplemental Table S1).

624 (87%) of the 716 human SBS-containing promoters that we identified at DBTSS overlap CpG islands, sequences known to be associated with active promoters (23, 24) (supplemental Table S1). In general, the promoters we found were GC-rich, and we could not recognize any significant TATA box sequence. TATA box consensus sequences could be observed in only 19 (5%) of the 355 examined promoters, suggesting that they are not a characteristic feature of SBS-containing gene promoters. Summarized in supplemental Table S1 are various features of the identified SBS and promoters, including identification number of the human promoters, gene name and symbol, position of the SBS in human and mouse promoters, coordinates of SBS on human chromosomes, motif orientation, association with CpG islands, sequence of the SBS motif with the associated 5’ submotif, alignment of the mouse and human motifs, and motif conservation in rat and dog genomes.

**Occupancy of Endogenous Genes by hStaf/ZNF143—**To determine whether hStaf/ZNF143 actually binds to the conserved motifs, we tested by ChIP assays 430 of the sites identified in silico and contained in 295 different promoters (listed in supplemental Table 1; Fig. 4A). The 295 promoters represent a statistically valid sample because they include 107 (56%) promoters with multiple sites (multiple site promoter (MSP)) and 188 (25%) with unique sites (unique site promoter (USP)). The tested MSP corresponded essentially to genes with an attributed function. The tested USP were extracted with random choice from the results of the different queries at DBTSS. To obtain statistically valid data, the sample of tested promoters containing SBS with and without the ACTACAN motif was directly proportional to the outcome of the various interrogation results. ChIP analysis was performed in HeLa cells with antibodies directed against hStaf/ZNF143. The recovered DNA was analyzed by semi-quantitative PCR with primer pairs spanning promoters with either single or multiple SBS. Each DNA sequence was amplified in parallel reactions with each of the three DNA templates purified from the following: 1) anti-hStaf/ZNF143 ChIP; 2) control ChIP; and 3) input chromatin. For each PCR, we tested two dilutions of DNA immunoprecipitated with anti-hStaf/ZNF143 (Fig. 4, B–G, lanes 1 and 2; supplementen-

---

**TABLE 2**

Identification of Staf-binding sites by searching transcription factor-binding sites at DBTSS

| Query name | Sequence | New human/mouse hits |
|------------|----------|----------------------|
| LA1        | A1 (YCCCAANNRRC:NYYRC) | ACTACAN 12 |
| LB1        | B1 (---------R----------) | ACTACAN 0    |
| LC1        | C1 (---------R----------) | ACTACAN 6    |
| LD1        | D1 (---------G----------) | ACTACAN 1    |
| LE1        | E1 (---------R----------) | ACTACAN 1    |
| LF1        | F1 (---------Y----------) | ACTACAN 0    |
| LG1        | g1 (---------Y----------) | ACTACAN 1    |
| LH1        | h1 (---------G----------) | ACTACAN 0    |
| LI1        | i1 (---------Y----------) | ACTACAN 0    |
| LJa        | j1 (---------Y----------) | ACTACAN 0    |
| LK1        | k1 (---------R----------) | ACTACAN 0    |
| LA1–LKa    | (---------R----------) | ACTACAN 0    |

**TABLE 3**

Identification of Staf-binding sites by searching transcription factor-binding sites at DBTSS

| Query name | Sequence motif 1 | Sequence motif 2 | New human/mouse hits |
|------------|-----------------|-----------------|----------------------|
| LA1        | A1 (YCCCAANNRRC:NYYRC) | ACTACAN 12 | |
| LB1        | B1 (---------R----------) | ACTACAN 0 | |
| LC1        | C1 (---------R----------) | ACTACAN 6 | |
| LD1        | D1 (---------G----------) | ACTACAN 1 | |
| LE1        | E1 (---------R----------) | ACTACAN 1 | |
| LF1        | F1 (---------Y----------) | ACTACAN 0 | |
| LG1        | g1 (---------Y----------) | ACTACAN 1 | |
| LH1        | h1 (---------G----------) | ACTACAN 0 | |
| LI1        | i1 (---------Y----------) | ACTACAN 0 | |
| LJa        | j1 (---------Y----------) | ACTACAN 0 | |
| LK1        | k1 (---------R----------) | ACTACAN 0 | |
| LA1–LKa    | (---------R----------) | ACTACAN 0 | |

---

**hStaf/ZNF143-binding Sites in the Human Genome**

are conserved in the rat and dog genomes, with 273 (96.8%) occurring also in the mouse genome (supplemental Table S1). Collectively, our in silico analysis resulted in the identification of 1175 conserved putative SBS harbored in 938 promoters of mammalian protein coding genes (Fig. 2A). Considering the results of this phylogenetic footprinting, it is likely that a high proportion of the identified SBS represent target sites for hStaf/ZNF143.

**Characteristic Features of the Identified SBS and SBS-containing Promoters—**Two features characterize a transcription factor-binding site as follows: first, its position, and second, its orientation relative to the gene. Fig. 2B (see also supplemental Table S1) shows the number of human SBS, with assigned positions, plotted versus their position relative to the TSS. We found that 673 (62%) of the 1077 SBS with assigned positions occur within 200 bp upstream of the TSS. Similar values were obtained with the SBS identified in mouse promoters, consistent with the involvement of this motif in initiation of transcription. We observed that 32% of the human/mouse orthologous SBS are positioned at the same distance from the TSS but that a difference of up to 50 bp occurs in 47% of the cases. It is very likely that this variation is related to the lack of accuracy in determining the TSS of numerous genes.

Another feature of the SBS motif that we detected is its property to appear as multiple copies. Overall, 188 (20%) of the 939 mammalian promoters we have identified contained multiple SBS lying within 500 bp (Fig. 2C). The majority (150) contain two SBS, with three SBS being found in 31 promoters; four genes contain four SBS (CDKS (cyclin-dependent kinase 5), ENY2 (enhancer of yellow 2 homolog), NUDCD1 (NudC domain containing 1), and PRCC (papillary renal cell carcinoma)); two genes have five SBS (LIG1 (DNA ligase 1) and BANP (BTG3 associated nuclear protein)); one single gene represents a unique situation with six SBS (NCLN (nicalin homolog)). Fig. 3 displays the multiple sequence alignment of the nicalin homolog promoter region across the four mammalian genomes. The very high degree of sequence conservation of the six potential SBS appears clearly and stands out from the nonconserved flanking sequences.
tal Fig. S1) or no antibody (Fig. 4, B–G, lanes 3 and 4; supplemental Fig. S1). Also, serial dilutions of input material were analyzed to demonstrate that the PCR was quantitative within a linear range of amplification (Fig. 4, B–G, lanes 5–7; supplemental Fig. S1). The positive controls for the ChIP assay included genes previously reported to respond to hStaf/ZNF143, the tRNA^Sec^, synaptoprevin-like1, aldehyde reductase, and t-complex polypeptide 1 (1, 11, 15, 17). As expected (Fig. 4B), a specific signal of higher intensity than in the no antibody control was obtained with the DNA immunoprecipitated with anti-hStaf/ZNF143 (compare lanes 1 and 2 and 3 and 4 in Fig. 4B). In contrast, no specific signal could be obtained with the primer pairs P1 to P4 amplifying DNA sequences lying several kbp upstream of the tRNA^Sec^, U4 ATAC, GAPDH, and BLUB1B genes because these remote regions were not expected to interact with hStaf/ZNF143 (compare lanes 1 and 2 and 3 and 4 in Fig. 4C). Among the 295 PCR amplifications, 29 (9.8%) were close to the background (supplemental Table S1 and Fig. S1). This is illustrated in Fig. 4, E and G, where the assay was performed with promoters containing two (SKB1 and MRPS25) and one (ZNF345 and ZNF451) putative SBS. Of the remaining promoters that were tested, however, 100/107 MSP (93.4%) and 166/188 USP (88%) provided clear positive signals (supplemental Table S1 and Fig. S1). Typical results are shown for 10 promoters containing two to four SBS (VAPA, WASF1, TMEM59, CDKS, SPPIN1, BRD8, DNML1, HCFC1R1, NMT1, and MRPL24; Fig. 4D) and 10 promoters containing one SBS (MTR, KIF11, SCAMP3, MDH1, SMUG1, PUM1, FLJ1311, TAF6, LACE1, and MGC29463; Fig. 4F). We observed an identical percentage of positive results regardless of the sequence used for identifying the motifs. Furthermore, no bias was introduced whether promoters with or without the ACTACA N motif were employed. By inspecting the SBS sequences that overlap the M4 motif of Xie et al. (20) (Xie in supplemental Table S1), about 15% had the highly conserved C12 substituted by A, G, or T (compare sequence logos in Fig. 1, A and E). ChIP experiments with five of them (A12, GenBank accession numbers NM_001320 and NM_006185; G12, GenBank accession numbers NM_004748 and NM_012257; T12, GenBank accession number NM_005111; supplemental Fig. S1) concluded that sequence variation at position 12 did not influence the binding of hStaf/ZNF143. Taken together, we could confirm that 266 (90%) of the promoters tested did harbor true hStaf/ZNF143-binding sites, underpinning the robustness of the computational screens. Moreover, these findings strongly suggest that the hStaf/ZNF143-binding sites identified in silico constitute high prevalence bona fide direct targets of hStaf/ZNF143.
The SBS Motifs Are Sufficient to Direct Transcriptional Activation without Associated Basal Promoter Elements—Only 5% of the promoters identified in this study contain a putative TATA box. To assess whether the SBS alone has the capacity to drive transcription, we tested its ability to up-regulate a luciferase reporter gene in a context deprived of known basal promoter elements. Reporters were constructed that contained three identical copies of wild type (WT SBS) or mutant SBS (mut-1 SBS, mut-2 SBS; Fig. 5A) directly fused to the luciferase coding sequence. In these constructs, the SBS reside at −91, −149, and −182 bp upstream of the luciferase initiation codon. The effect of those same mutations was evaluated in a previous work but in a different context, with mut-1 leading to a moderately reduced ability to bind Staf, whereas mut-2 affected its binding much more severely (1). Transfection of the WT reporter into COS-7 cells resulted in efficient transcriptional activation (compare empty vector and WT SBS in Fig. 5B). In contrast, the reporters bearing mutations in the central (mut-1 SBS) or in the 5′ part (mut-2 SBS) of the motif resulted in 8- and 22-fold reduction of transcriptional levels, respectively (Fig. 5B, compare WT SBS, mut-1 SBS, and mut-2 SBS). Taken together, these results indicate that the SBS is sufficient to direct expression of a reporter gene in the absence of known basal promoter elements.

To further evaluate the role of the ACTACAN motif, the effects of mutations therein were evaluated on the activity of a luciferase reporter gene directed only by the SBS and ACTACAN motifs. We constructed three different reporters containing, directly fused to the luciferase coding sequence, two identical copies of the following: 1) the wild-type ACTACAN motif and WT SBS (WT ACTACAN and WT SBS); 2) the mutated ACTACAN and WT SBS (mut ACTACAN and WT SBS); 3) the mutated ACTACAN motif and mutated SBS (mut ACTACAN and mut SBS) (Fig. 5A). In these constructs, the SBS resides at −189 and −303 bp upstream of the luciferase translation initiation codon. The construct containing both mutated motifs (mut ACTACAN and mut SBS) was unable to drive transcription above the basal level of the empty vector (Fig. 5C). In contrast, the presence of the WT ACATCA associated with the WT SBS induced a 49-fold induction of the luciferase activity (Fig. 5C). Moreover, mutation of the ACTACAN motif alone reduced the
transcriptional activity to about 50% of the WT level (Fig. 5C). Collectively, these results demonstrated the functional importance of the ACTACAN motif associated with the SBS.

**DISCUSSION**

The hStaf/ZNF143 Transcription Factor-binding Site, One of the Most Widespread in the Human Genome—We have performed a large scale analysis to evaluate the binding of hStaf/ZNF143 to mammalian promoters containing the SBS consensus element and variants therein. Our data revealed that the protein binds to a strikingly large number of genes, suggesting a significant diversity in the ensuing transcriptional response. In total, we have defined a set of 1175 evolutionarily preserved hStaf/ZNF143-binding sites in mammalian genomes and distributed in 938 promoters. Direct experimental validation by ChIP on 295 promoters yielded 90% success, indicating that very few of the identified sites were false positives. Moreover, negative results with ChIP are not necessarily synonymous of protein absence on the promoter, this technique being dependent on factors such as accessibility of the antibody to its epitope (25). Importantly, the high rate of success demonstrates the robustness of the in silico approach applied here.

Previously, hStaf/ZNF143-binding sites were identified by studying individual genes. At the onset of this work, only seven promoters of protein coding genes were recognized as hStaf/ZNF143 targets (11–17). Among the 938 promoters identified in this study, 716 were obtained by queries at DBTSS (version 4), which contains only 8793 human promoters. Extrapolation to the 25,000 human genes (26) revealed a minimum estimate of 2500 SBS in 2000 promoters of protein coding genes. This value is very similar to the number of Sp1-binding sites in human genes. This value is very similar to the number of Sp1-binding sites in human genes. This value is very similar to the number of Sp1-binding sites in human genes. This value is very similar to the number of Sp1-binding sites in human genes.

In total, we have defined a set of 1175 evolutionarily preserved hStaf/ZNF143-binding sites in mammalian genomes and distributed in 938 promoters. Direct experimental validation by ChIP on 295 promoters yielded 90% success, indicating that very few of the identified sites were false positives. Moreover, negative results with ChIP are not necessarily synonymous of protein absence on the promoter, this technique being dependent on factors such as accessibility of the antibody to its epitope (25). Importantly, the high rate of success demonstrates the robustness of the in silico approach applied here.

Previously, hStaf/ZNF143-binding sites were identified by studying individual genes. At the onset of this work, only seven promoters of protein coding genes were recognized as hStaf/ZNF143 targets (11–17). Among the 938 promoters identified in this study, 716 were obtained by queries at DBTSS (version 4), which contains only 8793 human promoters. Extrapolation to the 25,000 human genes (26) revealed a minimum estimate of 2500 SBS in 2000 promoters of protein coding genes. This value is very similar to the number of Sp1-binding sites in human genes. This value is very similar to the number of Sp1-binding sites in human genes. This value is very similar to the number of Sp1-binding sites in human genes.
It may well be that this sub-motif serves as a binding site for an unknown transcription factor. It might also induce a particular structure to the DNA that will optimize the binding of hStaf/ZNF143 to its cognate sequence. The ACTACAN sequence is particularly interesting in the light of a recent work describing the systematic discovery of motifs in human promoters by comparing several mammalian genomes (20). This promoter analysis yielded candidate motifs, including most of the previously known transcription factor-binding sites but also novel motifs. The top six of the discovered motifs and associated factors are M1, NRF-1; M2, MYC; M3, ELK-1; M4, unknown factor; M5, NFY; M6, Sp1 with a motif conservation score (MCS) of 107.8, 85.3, 80.4, 69.5, 64.6, and 63.9, respectively. Motifs with high MCS are both highly conserved and frequently occurring in the promoters of protein genes. The MCS of the M4 motif ACTA\textsuperscript{Y-RNNN}CCC\textsuperscript{R} has conservation and occurrence frequency higher than that of Sp1 and NFY transcription factors. Our work established clearly that ACTA\textsuperscript{YRN} indeed corresponds to the ACTACA\textsuperscript{N} sequence we identified and that NNCCCR represents SBS residues 1–6 (Fig. 1A, right panel). Xie et al. (20) did not report residues 7–18 of the SBS as a conserved motif. This is very likely because of the lower sequence conservation in the 3' part of the SBS. Considering that 42% of the identified SBS lacked the ACTACAN submotif, the global MCS for the hStaf/ZNF143-binding site (with and without ACTACAN) is higher than the MCS of M4. This strongly suggests that the hStaf/ZNF143-binding site represents one of the most frequently occurring transcription factor-binding sites in the human genome.

**Identified SBS Are Located Upstream TAF Sites**—Promoters with SBS showed high enrichment in CpG islands. The percentage of CpG-associated promoters (87%) is equivalent to the percentage (88%) of CpG associated with active pro-
hStaf/ZNF143-binding Sites in the Human Genome

motors (29). Notably, we did not find the TATA box to be significantly present in SBS-associated promoters, consistent with the TATA box being not a general promoter element in mammalian genes (29). Although the presence of a hStaf/ZNF143-binding site does not necessarily imply a direct effect of the protein on the expression of the gene, it is difficult to imagine that hundreds of hStaf/ZNF143 sites lying at the proximity of TSS could represent random and functionally meaningless occurrences. In this respect, we have shown that introduction of an SBS is sufficient to up-regulate transcription of a luciferase reporter gene, suggesting that hStaf/ZNF143 can directly or indirectly recruit the transcription machinery.

We examined the results of Kim et al. (29) who performed genome-wide location analysis experiments to identify gene promoters associated with TAF1, the largest subunit of TFIID. They concluded that a total of 9328 nonredundant gene promoters are occupied by TAF1. We found that 346 hStaf/ZNF143-bound promoters also harbor TAF1-binding sites. Comparison of the positions of SBS- and TAF1-binding sites in one given promoter established that 94.5% SBS are found upstream of TAF1 locations. Analysis of the distribution of the SBS and TAF1 sites revealed the interesting feature that SBS are essentially upstream of the transcription start site, whereas TAF sites are mostly found downstream of it (Fig. 2D). Moreover, the number of common promoter targets between TFIID and hStaf/ZNF143 far exceeds a random overlap between two similar groups, suggesting that hStaf/ZNF143 functions with TFIID in regulating gene expression. In this respect, we showed that SBS are sufficient to up-regulate a reporter gene on their own, and it is tempting to speculate that hStaf/ZNF143 can directly recruit TFIID.

SBS and Bidirectional Promoters—It is well established that bidirectional promoters are relatively common in mammalian genomes, representing more than 10% of the genes whose TSS are separated by less than 1000 bp (30, 31). In 15% of the identified promoters (146 promoters distributed in 73 pairs), a common SBS is present in two promoters in a head-to-head orientation. Further studies are needed to elucidate the direct involvement of hStaf/ZNF143 in this particular gene organization.

Functional Classification of the hStaf/ZNF143 Target Genes—Among the identified promoters, 30% were associated with genes with unknown function. To determine whether hStaf/ZNF143 is specifically associated with a gene category with a particular function, we categorized the genes according to their functional annotation. As shown in Fig. 6, the identified genes are distributed along all the functional categories. The prominent classes are as follows: (i) DNA-binding proteins and transcription factors that represent 23% of the total; (ii) protein synthesis degradation turnover and modification (21% of the total). We note that many of the identified genes are indeed important for cell growth.

In summary, our ChIP-coupled comparative genomic analysis has delivered a first version of the human catalog of hStaf/ZNF143 regulatory motifs in gene promoters. This work provides a convincing demonstration that performing combined 
in silico and experimental studies is absolutely instrumental in analyzing gene promoters.

Acknowledgments—We are grateful to F. Jossinet for statistical analysis and A. Schweigert for excellent technical assistance.

REFERENCES

1. Schuster, C., Myslinski, E., Krol, A., and Carbon, P. (1995) EMBO J. 14, 3777–3787
2. Schaub, M., Myslinski, E., Schuster, C., Krol, A., and Carbon, P. (1997) EMBO J. 16, 173–181
3. Myslinski, E., Ame, J. C., Krol, A., and Carbon, P. (2001) Nucleic Acids Res. 29, 2502–2509
4. Schaub, M., Krol, A., and Carbon, P. (1999) J. Biol. Chem. 274, 24241–24249
5. Schaub, M., Myslinski, E., Krol, A., and Carbon, P. (1999) J. Biol. Chem. 274, 25042–25050
6. Ragoussis, J., Senger, G., Mockridge, I., Sanseau, P., Ruddy, S., Dudley, K., Sheer, D., and Trowsdale, J. (1992) Genomics 14, 673–679
7. Tommerup, N., and Vissing, H. (1995) Genomics 27, 259–264
8. Myslinski, E., Krol, A., and Carbon, P. (1998) J. Biol. Chem. 273, 21998–22006
9. Myslinski, E., Krol, A., and Carbon, P. (1992) Nucleic Acids Res. 20, 203–209
10. Schuster, C., Krol, A., and Carbon, P. (1998) Mol. Cell. Biol. 18, 2650–2658
11. Kubota, H., Yokota, S., Yanagi, H., and Yura, T. (2000) J. Biol. Chem. 275, 28641–28648
12. Mach, C. M., Hargrove, B. W., and Kunkel, G. R. (2002) J. Biol. Chem. 277, 4883–4888
13. Saur, D., Seidler, B., Paehge, H., Schusdziarra, V., and Allescher, H. D. (2002) J. Biol. Chem. 277, 25798–25814
14. Grossman, C. E., Qian, Y., Banki, K., and Perl, A. (2004) J. Biol. Chem. 279, 12190–12205
15. Barski, O. A., Papusha, V. Z., Kunkel, G. R., and Gabbay, K. H. (2004) Genomics 83, 119–129
16. Ishiguchi, H., Izumi, H., Torigoe, T., Yoshida, Y., Kubota, H., Tsuji, S., and Kohno, K. (2004) Int. J. Cancer 111, 900–909
17. Di Leva, F., Ferrante, M. I., Demarchi, F., Caravelli, A., Matarazzo, M. R., Giaccia, M., D’Urso, M., D’Esposito, M., and Franze, A. (2004) J. Biol. Chem. 279, 7734–7739
18. Suzuki, Y., Yamashita, R., Nakai, K., and Sugano, S. (2002) Nucleic Acids Res. 30, 328–331
19. Yamashita, R., Suzuki, Y., Wakaguri, H., Tsuritani, K., Nakai, K., and Sugano, S. (2006) Nucleic Acids Res. 34, D86–D89
20. Xie, X., Lu, J., Kubokas, E. J., Golub, T. R., Mootha, V., Lindblad-Toh, K., Lander, E. S., and Kellis, M. (2005) Nature 434, 338–345
21. Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004) Genome Res. 14, 1188–1190
22. Weinmann, A. S., and Farnham, P. J. (2002) Methods (Orlando) 26, 37–47
23. Antequera, F. (2003) Cell. Mol. Life Sci. 60, 1647–1658
24. Caiafa, P., and Zampieri, M. (2005) J. Cell. Biochem. 94, 257–265
25. Blais, A., and Dunlacht, B. D. (2005) Genes Dev. 19, 1499–1511
26. International Human Genome Sequencing Consortium (2004) Nature 431, 931–945
27. Cawley, S., Bekirranov, S., Ng H. H., Kapranov, P., Sekinger, E. A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A. J., Wheeler, R., Wong, B., Drenkow, J., Yamanaka, M., Patel, S., Brubaker, S., Tammana, H., Helt, G., Struhl, K., and Gingeras, T. R. (2004) Cell 116, 499–509
28. Myslinski, E., Krol, A., and Carbon, P. (2004) Gene (Amst.) 330, 149–158
29. Kim, T. H., Herrera, L., Oso, Z., Ma, C., Singer, M. A., Richmond, T. A., Wu, Y., Green, R. D., and Ren, B. (2005) Nature 436, 876–880
30. Adachi, N., and Lieber, M. R. (2002) Cell 109, 807–809
31. Trinklein, N. D., Aldred, S. F., Hartman, S. J., Schroeder, D. I., Otillar, R. P., and Myers, R. M. (2004) Genome Res. 14, 62–66