Transcription factor binding and modified histones in human bidirectional promoters

Jane M. Lin, Patrick J. Collins, Nathan D. Trinklein, Yutao Fu, Hualin Xi, Richard M. Myers and Zhiping Weng

Genome Res. 2007 17: 818-827
Access the most recent version at doi:10.1101/gr.5623407

Letter

Transcription factor binding and modified histones in human bidirectional promoters

Jane M. Lin,[1] Patrick J. Collins,[2] Nathan D. Trinklein,[2,4] Yutao Fu,[3] Hualin Xi,[3] Richard M. Myers,[2] and Zhiping Weng[1,3,5]

[1]*Department of Biomedical Engineering, Boston University, Boston, Massachusetts, 02215, USA;* [2]*Department of Genetics, Stanford University, School of Medicine, Stanford, California 94305-5120, USA;* [3]*Program in Bioinformatics and Systems Biology, Boston University, Boston, Massachusetts, 02215, USA*

Bidirectional promoters have received considerable attention because of their ability to regulate two downstream genes (divergent genes). They are also highly abundant, directing the transcription of ~11% of genes in the human genome. We categorized the presence of DNA sequence motifs, binding of transcription factors, and modified histones as overrepresented, shared, or underrepresented in bidirectional promoters with respect to unidirectional promoters. We found that a small set of motifs, including GABPA, MYC, E2F1, E2F4, NRF-1, CCAAT, YY1, and ACTACAnnTCC are overrepresented in bidirectional promoters, while the majority (73%) of known vertebrate motifs are underrepresented. We performed chromatin-immunoprecipitation (ChIP), followed by quantitative PCR for GABPA, on 118 regions in the human genome and showed that it binds to bidirectional promoters more frequently than unidirectional promoters, and its position-specific scoring matrix is highly predictive of binding. Signatures of active transcription, such as occupancy of RNA polymerase II and the modified histones H3K4me2, H3K4me3, and H3ac, are overrepresented in regions around bidirectional promoters, suggesting that a higher fraction of divergent genes are transcribed in a given cell than the fraction of other genes. Accordingly, analysis of whole-genome microarray data indicates that 68% of divergent genes are transcribed compared with 44% of all human genes. By combining the analysis of publicly available ENCODE data and a detailed study of GABPA, we survey bidirectional promoters with breadth and depth, leading to biological insights concerning their motif composition and bidirectional regulatory mode.

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

Mammalian genomes are highly complex, with neighboring genes arranged in divergent, convergent, tandem, anti-sense, and interleaving fashions (Carninci et al. 2005, 2006; Kapranov et al. 2005; Carninci 2006; Engstrom et al. 2006). Despite the vast genomic space, a substantial portion of human genes (~11%) are arranged in a divergent, head-to-head fashion and controlled by bidirectional promoters (Trinklein et al. 2004). We define a bidirectional promoter as an intergenic region that is <1 kb long, and flanked by the transcription start site (TSS) of a plus-strand gene on one side and the TSS of a minus-strand gene on the other. This abundance has been observed across several mammalian genomes (Adachi and Lieber 2002; Koyanagi et al. 2005), suggesting that there is evolutionary pressure for conserving this type of gene-pair structure. Some divergent genes are related by function, in particular DNA repair (Adachi and Lieber 2002; Trinklein et al. 2004), and could be coregulated in a way that takes advantage of their paired arrangement. Indeed, the expression patterns of divergent gene pairs are more correlated than those of randomly paired genes (Trinklein et al. 2004).

Investigators have examined the sequences of bidirectional promoters for clues as to how they regulate both downstream genes. For example, most bidirectional promoters lack TATA boxes and are both GC-rich and enriched in CpG islands (Adachi and Lieber 2002; Takai and Jones 2004; Trinklein et al. 2004). They also display a mirror sequence composition, such that Gs and Ts dominate on one side of the midpoint, while Cs and As dominate on the other side (Engstrom et al. 2006). How these general features influence the binding of transcription factors (TFs) to bidirectional promoters is unknown. While numerous studies have summarized the composition of sequence motifs in unidirectional promoters (Ettwiller et al. 2005; Xie et al. 2005), none have addressed bidirectional promoters specifically. TF-binding sites of individual bidirectional promoters have been experimentally identified (Carter and Avadhani 1994; Bush et al. 2003; Patton et al. 2005), but the results are difficult to generalize. In this study, we formulate a three-category conceptual framework to classify sequence motifs in bidirectional promoters computationally as overrepresented, shared, or underrepresented (Fig. 1).

We took advantage of the immense repository of ChIP-chip target lists that were made public by The ENCODE Project Consortium to evaluate the biological relevance of our computational categorization. We analyzed 32 ChIP-chip experiments on 15 factors, including nine TFs, four types of modified histones, RNA polymerase II (POLR2A), and TATA-associated factor 1 (TAF1), and classified them as overrepresented, shared, or underrepresented. We complemented this survey by performing an in-depth case study of one overrepresented TF, GABPA, which has consensus binding sites in many bidirectional promoters. We performed GABPA ChIP-qPCR on 118 regions in the human genome and experimentally confirmed that the occupancy of

[4]Present address: SwitchGear Genomics, Menlo Park, CA 94025, USA.
[5]Corresponding author.
E-mail zhiping@bu.edu; fax (617) 353-6766.
Article is online at http://www.genome.org/cgi/doi/10.1101/gr.5623407. Freely available online through the *Genome Research* Open Access option.

17:818–827 ©2007 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/07; www.genome.org

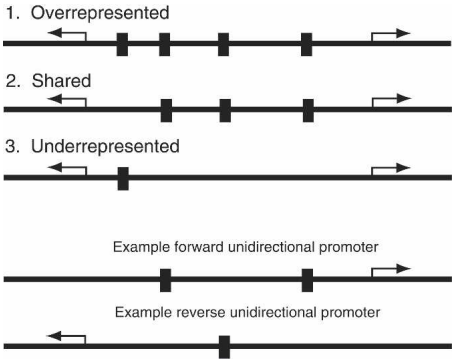


Figure 1. Three types of motif representation in bidirectional promoters. Motifs in the overrepresented category occur more often (e.g., four times) than the sum of occurrences in two unidirectional promoters (e.g., 2 + 1 = 3 times). Motifs in the shared category occur as often (e.g., three times) as the sum of occurrences in two unidirectional promoters. Motifs in the underrepresented category occur less often (e.g., once) than the sum of occurrences in two unidirectional promoters.

GABPA is overrepresented in bidirectional promoters. Furthermore, we found that regions bound by GABPA tend to have high-scoring (see Methods for scoring definitions) binding sites. In our previous study, we made a series of deletions in several bidirectional promoters and showed that most of these contain a shared fragment necessary for full promoter activity in both directions in a transient transfection assay system (Trinklein et al. 2004). In the current study, we further mutated bases throughout the 30-bp fragment from one promoter and identified 12 bases overlapping the GABPA consensus that were essential for bidirectional promoter activity.

Because bidirectional promoters have the ability to activate two downstream genes concurrently, one of our goals was to determine whether there are sequence signatures that distinguish this mode of regulation from that of unidirectional promoters. In addition to identifying such signatures, we found evidence for overrepresentation for the binding of certain TFs and modified histones. We experimentally verified the binding of a transcription factor, GABPA, in bidirectional promoters. Detailed analysis reveal that the dominant mode of transcription in bidirectional promoters is to activate both downstream genes, and that this can be achieved with as little as a single TF binding site.

Results

Three types of TF motifs

A bidirectional promoter can be viewed simply as two unidirectional promoters sharing the same genomic space. Consequently, we would expect the occurrence of a motif in a bidirectional promoter to be, on average, equal to the sum of the occurrences of the motif in two unidirectional promoters. Systematic deviation from the sum would indicate exceptional motif composition in bidirectional promoters.

We define a three-category framework to describe motif representation in bidirectional promoters (Fig. 1): (1) Overrepresented: The number of binding sites in bidirectional promoters is larger on average than the sum of sites in two unidirectional promoters. (2) Shared: The number of binding sites is larger in bidirectional promoters than is the number of sites in a single unidirectional promoter, but smaller than the sum of sites in two unidirectional promoters. (3) Underrepresented: The number of binding sites in bidirectional promoters is smaller than the number of sites in a single unidirectional promoter.

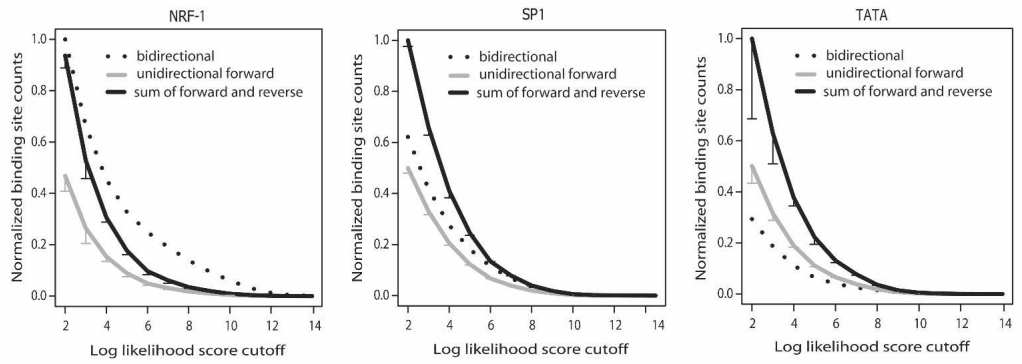
We hypothesize that overrepresented motifs correspond to TFs that are more likely to bind to bidirectional promoters than to unidirectional promoters, so much so that their binding-site abundance in bidirectional promoters exceeds expectation (sum of sites in two unidirectional promoters). In contrast, underrepresented motifs correspond to TFs that preferentially regulate unidirectional promoters. TFs of shared motifs show no preference, possibly corresponding to potent regulators, as they do not require augmentation to the number of binding sites for effective usage in bidirectional promoters, assuming multiple sites lead to a higher likelihood of binding.

Ab initio motif discovery in bidirectional promoters

We first wished to determine whether bidirectional promoters distinguish themselves from unidirectional promoters by housing a special set of motifs. We compiled two data sets of bidirectional promoters and one data set of unidirectional promoters for comparison (see Methods for details). The smaller, higher quality set of 376 bidirectional promoters was analyzed by the ab initio motif discovery algorithm MEME (Grundy et al. 1996), and the sequences from a larger set of 1304 bidirectional promoters were used to categorize motifs as overrepresented, shared, or underrepresented. We randomly sampled 1304 pairs of unidirectional promoters from a total of 13,205 in the human genome and matched them with the 1304 bidirectional promoters for CpG content and length to avoid selection bias. We required MEME to output 15 motifs, but several of these were reverse complements of each other or single nucleotide repeats, and thus, the list was reduced to seven unique motifs (Supplemental Fig. S1). Among these, we categorized five as overrepresented (NRF-1, CCAAT, GABPA, YY1, and ACTACAnnTCCC), one as shared (SP1), and one as underrepresented (Novel 1) (Fig. 2).

The increased presence of these motifs in bidirectional promoters is consistent with the overrepresentation of certain Gene Ontology (GO) categories among divergent genes (Boyle et al. 2004; Harris et al. 2004). For example, NRF-1 is a key regulator of nuclear genes encoding components of the mitochondrial transcription and replication machinery (Gopalakrishnan and Scarpulla 1995), and divergent genes are enriched in the GO term "mitochondrion" (GO:0005739, $P = 1.2 \times 10^{-9}$). Overexpression of YY1 has been shown to accelerate DNA repair (Oei and Shi 2001) and there is also an enrichment for the molecular function "response to DNA damage stimulus" (GO:0006974, $P = 8.2 \times 10^{-6}$) among divergent genes. The motif for GABPA (also called NRF-2) is also categorized as overrepresented, and like NRF-1, is implicated in the transcriptional regulation of several subunits of mitochondrial enzymes (Scarpulla 2006). We revisit GABPA in a case study further below.

Similarly, CCAAT-boxes (bound by NF-Y) are frequently found in TATA-less promoters (Mantovani 1999), and TATA boxes are underrepresented in bidirectional promoters (Trinklein et al. 2004). We also discovered a motif with the consensus ACTACAnnTCCC and classified it as overrepresented in bidirectional promoters. This motif was previously reported by Xie et al. with the consensus ACTAYRnnnCCCR and was ranked fourth of 174 motifs in terms of conservation across several mammalian promoters (Xie et al. 2005). Because the top three motifs in their study correspond to known transcription factors, this is actually



Overrepresented	Shared	Underrepresented
Ab Initio Motifs		
CCAAT[a] GABPA[a] ACTACAnnTCCC YY1 NRF-1[a]	SP1[a]	Novel 1
TRANSFAC Motifs		
47 (20 unique)	117 (67 unique)	440 (230 unique)
ENCODE ChIP-chip		
MYC[b] E2F1[a] E2F4[a] SP1 (Jurkat) SP3 (Jurkat) POLR2A[b] TAF1[c] H3ac[b,c] H3K4me2[b,c] H3K4me3[b,c]	cJun SP1[a] (HCT116) SP3[a] (HCT116)	STAT1[a,b] SP1 (K562) SP3 (K562) SMARCC1[c] SMARCC2[c] H4ac[b,c]

Figure 2. Enrichment of motifs, TF-binding, modified histones, and other nonsequence-specific factors in bidirectional promoters. The number of binding sites in bidirectional and sampled unidirectional promoters are normalized by the maximum number of sites and plotted against POSSUM log likelihood scores. Counts in the unidirectional reverse gene set are not shown for clarity because they are very similar to the binding-site counts in the unidirectional forward gene set (solid gray). The solid-black lines are the result of summing binding sites in the unidirectional forward and unidirectional reverse gene sets. Error bars indicate the standard error of the mean binding-site counts derived from 10 randomizations. [a]A TF that has a corresponding TRANSFAC motif placed in the same category; [b]a TF whose categorization is consistent among multiple ChIP experiments in different cell types or conditions; [c]a nonsequence-specific factor.

the highest-ranking novel motif in their set. Given the enrichment of this motif in bidirectional promoters and its strong evolutionary conservation across mammalian promoters, we predict that its cognate transcription factor plays an important role in regulating bidirectional promoters.

All five overrepresented motifs in bidirectional promoters are among the most conserved motifs in mammalian promoters at large: NRF-1, CCAAT, GABPA, YY1, and ACTACAnnTCCC were ranked by Xie et al. as No. 1, 5, 11, 10, and 4, respectively (Xie et al. 2005). Interestingly, we did not discover any novel motif that binds exclusively to bidirectional promoters.

TRANSFAC motif presence in bidirectional promoters

We used our motif classification framework to compare abundances of known vertebrate motifs from TRANSFAC (Wingender et al. 1996; Fu and Weng 2005) in bidirectional and unidirectional promoters. Of 604 motifs in the TRANSFAC vertebrate set, 47 are overrepresented, 117 are shared, and 440 are underrepresented. The aforementioned seven ab initio motifs and their corresponding TRANSFAC motifs are consistent in their categorizations, demonstrating the robustness of our approach and the high quality of the position-specific scoring matrices (PSSMs) discovered by MEME. To circumvent the redundancy in TRANSFAC, we calculated Pearson correlations between all PSSM pairs in each category using the malign algorithm (Haverty et al. 2004), setting a cutoff at 0.2, and obtaining 20, 67, and 230 unique motifs for each category, respectively. In both the nonunique and unique cases, the majority (440/604 = 73% and 230/317 = 73%) of TRANSFAC vertebrate motifs are underrepresented in bidirectional promoters (Fig. 2; Supplement 2). These results are consistent with a model that divergent genes are regulated by a limited set of transcription factors, despite the fact that they make up a substantial portion of human genes.

Positional preference of motifs in bidirectional promoters

Approximately 28% of known motifs show significant positional preference relative to the transcription start sites (TSSs) of human genes (Xie et al. 2005). Because bidirectional promoters have

varying lengths, we assessed the positional preferences of the ab initio motifs on normalized promoter lengths. We plot the positional-specific histogram relative to the length of the bidirectional promoter in the column labeled "Relative" (Supplemental Fig. S2), where the number 0 is closest to the TSS of the reverse-strand gene, and the number 1 is closest to the TSS of the forward-strand gene. The positional-specific histograms in absolute bases are also aligned to the reverse-strand gene or the forward-strand gene. The CCAAT box has a bimodal distribution, with peaks at 0.3 and 0.7, corresponding to symmetric placement on the bidirectional promoter around one-third of the way in from each TSS. GABPA has a concave-shaped distribution, peaking close to the flanking TSSs, which is consistent with its position specificity of −23 in unidirectional promoters. Novel 1 appears to have a multipeak distribution. ACTACAnnTCCC, NRF-1, and SP1 display no position specificity, despite showing specificity in unidirectional promoters at −89, −62, and −63, respectively (Xie et al. 2005). One explanation is that the consensus sequences for NRF-1 and SP1 are composed almost entirely of Gs and Cs; thus, their binding sites are frequently found throughout the GC-rich bidirectional promoters. In general, all seven motifs exhibit less positional specificity in bidirectional promoters than in unidirectional promoters, consistent with a model that some of these motif sites are involved in regulating both of the divergent genes.

Correlated transcription profiles of divergent genes

Several studies reported that divergent genes have more correlated expression patterns than randomly paired genes (Adachi and Lieber 2002; Trinklein et al. 2004; Li et al. 2006). As described above, we observed a small set of well-studied motifs to be enriched in bidirectional promoters; thus, we wished to examine whether divergent genes as a group had correlated expression patterns. We computed the Pearson correlation coefficients between the expression profiles (Su et al. 2004) of divergent genes from different bidirectional promoters and indeed observed a significant upward shift in their distribution with respect to that of randomly paired genes from unidirectional promoters (Fig. 3; $P < 2.2 \times 10^{-16}$ by the Wilcoxon rank sum test). Divergent gene pairs from the same bidirectional promoters are even more correlated ($P = 1.2 \times 10^{-11}$ compared with mismatched divergent genes and $P < 2.2 \times 10^{-16}$ compared with genes of unidirectional promoters, by the Wilcoxon rank sum test). Setting cutoffs at the fifth and 95th percentiles of the background distribution defined as that of unidirectional promoters, 27% of divergent gene pairs were coregulated, and 1.7% were antiregulated. These figures are close to the 17% and 1.8% we reported earlier, in which we used a different set of microarray data (Trinklein et al. 2004).

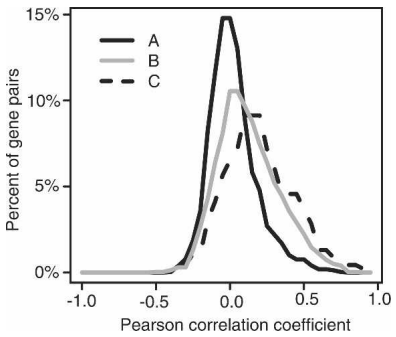


Figure 3. The distribution of Pearson correlation coefficients for (*A*) randomly paired genes of unidirectional promoters, (*B*) randomly paired divergent genes of different bidirectional promoters, and (*C*) divergent gene pairs of the same bidirectional promoters. Divergent genes from the same bidirectional promoters are more correlated in expression than randomly paired genes of unidirectional promoters ($P = 2.2 \times 10^{-16}$ by the Wilcoxon rank sum test) and randomly paired genes of different bidirectional promoters ($P = 1.21 \times 10^{-11}$).

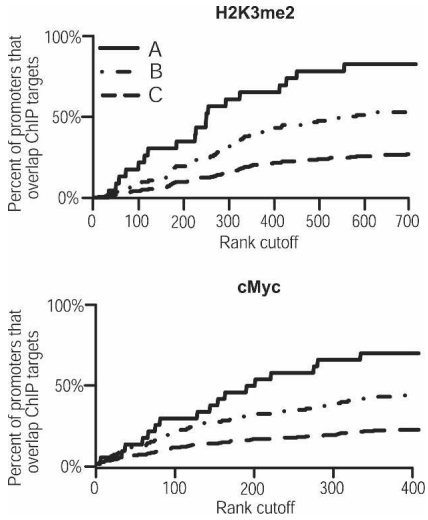


Figure 4. The fraction of promoters that overlap at least 1 bp with a set of ranked ChIP-chip target lists is plotted on the *y*-axis (see Methods for details). (A) Percent overlap of bidirectional promoters; (B) twice the percent overlap of unidirectional promoters; (C) percent overlap of unidirectional promoters. Both H3K4me3 and MYC overlap more with bidirectional promoters than twice the overlap of unidirectional promoters.

Bidirectional promoters and ChIP-chip experiments in the ENCODE regions

Taking advantage of publicly available ENCODE ChIP-chip data sets, we analyzed the binding preferences of seven sequence-specific factors, four sequence nonspecific factors, and four types of modified histones in a specified set of regions that make up 1% of the human genome (The ENCODE Project Consortium 2004). We analyzed the target lists of the ChIP-chip experiments reported at a 10% false discovery rate cutoff. By comparing the percent overlap (defined as within 500 bp) of these target lists in ENCODE bidirectional promoters with twice the percent overlap in ENCODE unidirectional promoters, we again categorized these factors and modified histones into overrepresented, shared, or underrepresented in bidirectional promoters (see Methods; Fig. 2).

The sequence-specific factors we analyzed included JUN, MYC, E2F1, E2F4, SP1, SP3, and STAT1. We found that MYC, E2F1, and E2F4 were overrepresented, as the percent overlap of their target lists with bidirectional promoters was more than twice that with unidirectional promoters at every target-list rank cutoff (Fig. 4). Our motif analysis above indicates that SP1 and SP3 were shared, but their ChIP data indicates that they preferentially bind to bidirectional promoters in Jurkat cells, show no preference in HCT116 cells, and show preference for unidirectional promoters in K562 cells. Careful examination reveals that this cell-line specificity is slight (Supplement 3), consistent with the classification of SP1 and SP3 as shared motifs. In summary,

bidirectional promoter binding in living cells for a majority of TFs (E2F1, E2F4, SP1, SP3, and STAT1) are classified in the same way as their canonical sequence motifs.

Among the four nonsequence-specific factors, TAF1 and POLR2A are overrepresented, while SMARCC1 and SMARCC2 are underrepresented in bidirectional promoters. Modified histones H3ac, H3K4me2, and H3K4me3 are overrepresented, while H4ac is underrepresented (Fig. 2). TAF1, POLR2A, and modified histones generally occur near the TSS, which might bias our analysis, but our method takes this into account by comparing percent overlap in bidirectional promoters with twice the percent overlap in unidirectional promoters. We show that these factors and modified histones occur in bidirectional promoters above and beyond the expectation at all rank cutoffs (Fig. 4). Because the presence of H3ac and H3K4me3 are powerful predictors of active transcription (Allfrey et al. 1964; Grant 2001; Bernstein et al. 2005), the overrepresentation of these modified histones and the TAF1 and POLR2A binding suggest that the proportion of transcribed divergent genes is larger than that of other genes.

We also analyzed the gene expression data on four cell lines (GM06990, HFL1, HeLa, and MOLT4) (Koch et al. 2007). We used the Present calls by Affymetrix MAS5.0 software to designate transcribed genes. Sixty-eight percent of divergent genes in ENCODE regions were consistently transcribed across these four cell lines, whereas only 45% of all human genes were consistently transcribed. This is consistent with the aforementioned results on modified histones and TAF1 and POLR2A binding.

The binding levels of modified histones around bidirectional promoters in the ENCODE regions

Levels of modified histones H3K4me2, H3K4me3, and H3ac have been shown to correlate positively with the level of transcription (The ENCODE Project Consortium 2007). We examined the average levels of these modified histones around individual bidirectional promoters by analyzing ChIP-chip experiments conducted on HeLa cells (Supplement 3). We observed that the ChIP intensities tended to be diminished throughout the bidirectional promoter region, but elevated immediately downstream of the flanking TSSs (Supplement 3; Supplemental Table S1). The average downstream ChIP intensities of all 46 high-quality divergent genes in the ENCODE regions were at similar levels as those of actively transcribed genes and much higher than those of all genes (Fig. 5).

To summarize the patterns of modified histones, we discretized the ChIP intensities as "elevated" or "diminished" in three regions: downstream of the reverse gene, throughout the bidirectional promoter, and downstream of the forward gene (see Methods). The most prominent histone modification pattern (32%) consisted of diminished signals in the bidirectional promoter and elevated signal downstream of both TSSs (labeled pattern 1 in Fig. 6). Smaller portions had diminished modification signals in the promoter and elevated signal downstream of either the forward TSS (9%; labeled 2 in Fig. 6) or the reverse TSS (15%; labeled 3 in Fig. 6). These patterns were most pronounced in H3K4me2 and H3ac (Fig. 6B,D).

A case study of GABPA

GABPA binds to a majority of bidirectional promoters

We computationally predicted GABPA-binding sites in the human genome and sampled 118 sites randomly from three types of genomic regions for testing: bidirectional promoters, unidirectional promoters, and nonpromoters. We further characterized the binding sites in each of these regions as containing high-scoring motifs, medium-scoring motifs, low-scoring motifs, or no motifs (see Methods for definition of motif scores). We then performed ChIP with an antibody recognizing GABPA and assayed the enrichment of precipitated fragments by quantitative real-time PCR (qPCR) in Jurkat and K562 cells (data available in Supplement 4). Figure 7 shows the breakdown of GABPA-bound fragments in the four motif categories at the fivefold enrichment cutoff. Our results show that GABPA binds to the majority (83%) of tested bidirectional promoters with high-scoring motifs (compared with 51% in unidirectional promoters and 14% in nonpromoters). From these data, we can estimate the proportion of GABPA-bound bidirectional promoters in the human genome as follows. GABPA binds to 83% of bidirectional promoters with high-scoring motifs and 21% of bidirectional promoters have at least one high-scoring motif, thus 17% (= 83%*21%) of bidirectional promoters have high-scoring motifs AND are bound by GABPA. Adding this to the estimates for medium- (16%) and low- (24%) scoring motifs gives a total of 57% of bidirectional promoters bound by GABPA. In contrast, it is estimated that only 7% of unidirectional promoters are bound by GABPA. This genome-wide estimate agrees well with our results in the ENCODE regions, for which we tested 16 of the 23 bidirectional promoters and found that nine (56%) were bound by GABPA in two cell lines and 12 (75%) were bound by GABPA in at least one cell line. Thus, GABPA can bind to a majority of bidirectional promoters, and its binding frequency is overrepresented in bidirectional promoters compared with unidirectional promoters. This suggests that GABPA is a major regulator of bidirectional transcription.

GABPA motif presence correlates with binding in living cells

The higher the motif score is for a site, the higher the likelihood that the site is occupied by GABPA. In bidirectional promoters, 44% of low-scoring sites were bound by GABPA compared with 68% of medium-scoring and 83% of high-scoring sites. Similarly,

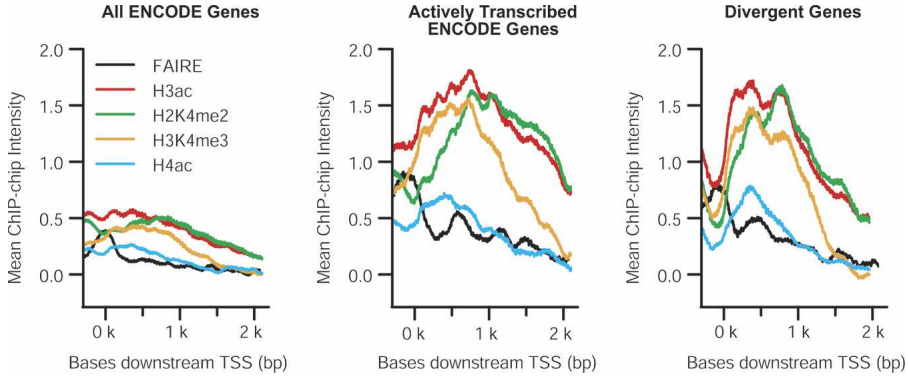


Figure 5. Histone modification ChIP intensities downstream of ENCODE genes. Average log2 ChIP intensity of several modified histones and formaldehyde-assisted isolation of regulatory elements (FAIRE) in all genes (Giresi et al. 2007), actively transcribed genes, and 46 divergent genes. Divergent genes have elevated histone modification signals downstream of their TSS, similar to actively transcribed genes.

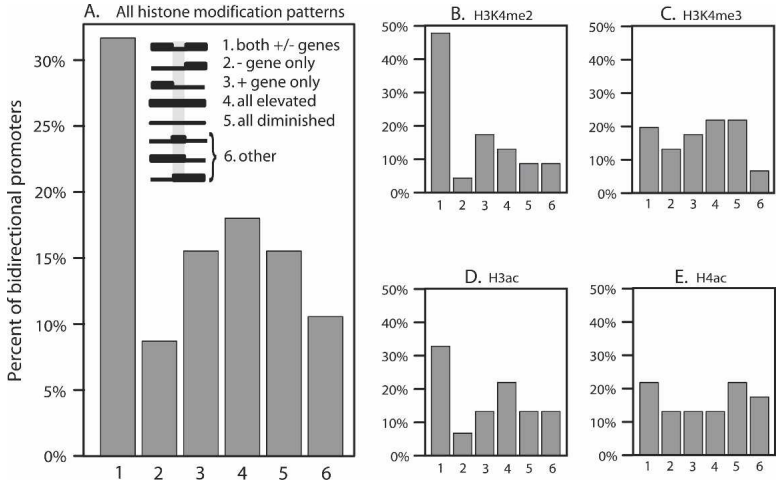


Figure 6. Discretized histone modification patterns are not uniformly distributed. The most prominent pattern is labeled 1, where ChIP intensities are elevated downstream of both genes, but diminished throughout the bidirectional promoter. Also prominent are patterns 2 and 3, where the ChIP intensities are just elevated in one downstream gene (the genomic forward direction or reverse direction, respectively). The pattern labeled 4 represents the case where ChIP intensities are elevated in the bidirectional promoter as well as both downstream genes, and 5 represents no elevation throughout the region. The number 6 represents three remaining patterns: elevated ChIP intensities in the bidirectional promoter but diminished in both downstream promoters, elevated ChIP intensities in the bidirectional promoter and downstream of the forward gene, but diminished in the reverse gene, and elevated ChIP intensities in the bidirectional promoter and downstream of the reverse gene, but diminished in the forward gene.

in unidirectional promoters, 0% low-, 42% medium-, and 51% high-scoring sites were bound by GABPA. All regions that we tested with no sites showed no binding, suggesting that the GABPA consensus sequence is necessary for binding in living cells. This appears to be unique to GABPA; other sequence-specific factors have been reported to bind many regions that do not contain a match for their canonical sequence motif (Bieda et al. 2006). GABPA could be driven to bind bidirectional promoters because a larger proportion of bidirectional promoters contain medium- and high-scoring binding sites (Supplemental Fig. S3A), and the number of binding sites is also greater (Supplemental Fig. S3B) compared with unidirectional promoters.

Eighty-six percent of the fragments we tested by ChIP contained more than one GABPA site. The fragments with the greatest enrichment usually contained one high-scoring site surrounded by several low-scoring sites. We sought to determine which characteristic—the score of the best binding site or the total number of binding sites—better predicted ChIP enrichment. We set the score threshold to the low level (log-likelihood score ≥ 3) and counted the total number of sites in the ChIP-identified fragment. The largest number of sites was seven, in a 185-bp long bidirectional promoter, while the average was three. The correlation between the total site count and ChIP enrichment (correlation coefficient = 0.40) is stronger than that between the highest site score and ChIP enrichment (correlation coefficient = 0.29), suggesting that having multiple sites is a better predictor of binding than having a single high-scoring GABPA site. In fact, these two characteristics frequently coincide; thus, it is possible that low-scoring sites help guide sequence-specific factors to high-scoring sites.

GABPA–binding sites are often required for bidirectional promoter activity

In a previous study, we systematically truncated and tested nine bidirectional promoters for activity in both directions by using a luciferase reporter assay (Trinklein et al. 2004). In the current study, we reanalyzed the truncated fragments by mapping binding sites of the aforementioned ab initio motifs in these nine promoters and found that eight of them contain at least one GABPA site that appears to be necessary for their bidirectional transcriptional activity. We define a binding site as necessary if the promoter activity was diminished in both directions after deletion of a fragment containing that site (Supplemental Fig. S4). The large number of necessary GABPA sites also supports its role as a major regulator of bidirectional promoters. However, in six of nine cases deletions of portions of the bidirectional promoter that did not have a GABPA site also showed diminished activity, indicating that there are other factors necessary for bidirectional promoter activity. For the case in Supplemental Figure S4A, a 30-bp region containing a single GABPA site was sufficient to drive transcription in both directions. Mutagenesis results further confirmed that 12 bases beginning at position 7 and overlapping the GABPA consensus CCGGAARYR are essential for bidirectional promoter activity (Fig. 8). The footprint of GABPA may extend beyond the 9-bp consensus sequence, as an additional three bases were found to be important for transcription (Mutation 9 in Fig. 8). Mutations 14 and 1 also led to decreased promoter activity; however, they are at the ends of the fragment and the decrease in activity occurs in only one direction; thus, they most likely correspond to the transcription start sites of the forward and reverse genes, respectively.

Discussion

We performed a comprehensive computational analysis of sequence motifs in bidirectional promoters, categorizing all 604 vertebrate motifs in the TRANSFAC database. To test predictions

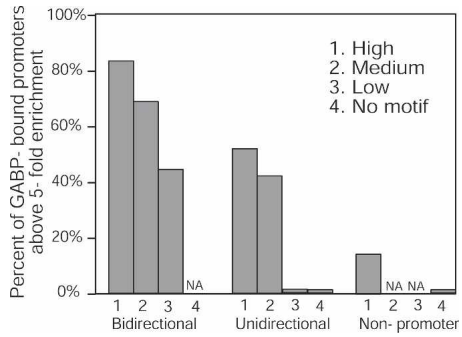


Figure 7. GABPA ChIP in three region categories and four motif score cutoffs. GABPA binds preferentially to bidirectional promoters over unidirectional promoters and over nonpromoter regions at all motif score cutoffs. Regions containing high-scoring binding sites are more likely to bind GABPA than those containing only medium or low-scoring sites. NA indicates that no regions in this category were tested.

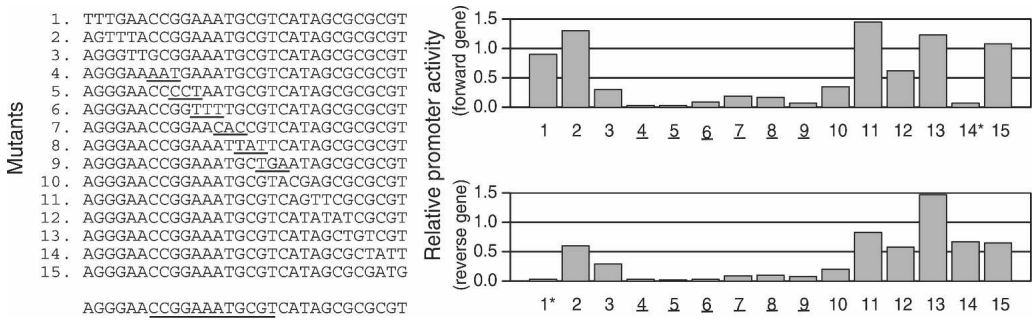


Figure 8. Mutagenesis of a 30-bp bidirectional fragment. Bases essential for promoter activity in each direction are underlined. Transversion mutations were introduced three bases at a time and shifted two bases for a total of 15 mutants. Mutation of a 12-bp region annihilates promoter activity in both directions. The consensus CCGGAARYR is recognized by the transcription factor GABPA. Asterisks (*) indicate position of transcription start sites.

made by this analysis, we used a large body of ChIP-chip data generated by the ENCODE Consortium, and also conducted detailed ChIP-chip and ChIP-qPCR experiments with an additional sequence-specific factor, GABPA. From this combination of predictions and experiments, we demonstrated the regulatory effect of GABPA on bidirectional promoters and identified an additional small set of sequence-specific transcription factors that appear to regulate bidirectional promoters preferentially. Interestingly, we did not uncover any motif exclusive to bidirectional promoters.

Our sequence analysis has been focused on the bidirectional promoters and not the divergent genes or regions further downstream. We intentionally excluded divergent genes that overlap exon sequences, as the promoters of these two genes do not overlap (they are in each other's genic space). Carninci et al. (2006) showed that many annotated nonoverlapping transcripts could also form overlapping transcripts. In these cases, each gene has a range of possible TSSs that varies within a 120-bp window. Because more than 120 bp separates the majority of gene pairs in our data set, the contamination by antisense transcripts is limited and probably has negligible effects on our results because we are investigating bulk properties of bidirectional promoters. We also note that although there are examples of regulatory elements in downstream and intronic regions (Kabat et al. 2006) and some may be exclusive to bidirectional promoters, we did not look for motifs beyond 50 bp downstream of the flanking TSSs.

The recognition sequences for GABPA, MYC, E2F1, E2F4, NRF-1, NFY, and YY1, as well as ACTACAnnTCC, are among the most conserved motifs found in human promoters. While they are found in all types of human promoters and not specialized to bidirectional promoters per se, they are significantly overrepresented in bidirectional compared with unidirectional promoters. Conversely, most (73%) vertebrate transcription factor motifs are underrepresented in bidirectional promoters. Thus, bidirectional promoters are characterized by the presence of binding sites for a limited set of transcription factors, and the absence of binding sites for the vast majority of transcription factors. This unusual composition of *cis*-acting sequences in bidirectional promoters is likely to be one of the major reasons why divergent genes are generally and broadly in active transcription mode. Accordingly, analysis of expression microarray data indicates that 68% of divergent genes are transcribed compared with 44% of all human genes. Moreover, divergent genes from different bidirectional promoters have significantly correlated expression patterns, likely reflecting the restricted motif composition of all bidirectional promoters at large.

Histone marks H3K4me2, H3K4me3, and H3ac are strongly indicative of transcriptional activity and, until now, they had not been examined specifically around human bidirectional promoters. Our conclusions based on histone modification ChIP-chip target lists and raw ChIP intensities are consistent, and suggest that divergent genes tend to be more actively transcribed than genes of unidirectional promoters. Additionally, because the ChIP signals for all the histone marks are diminished throughout bidirectional promoters, it is likely that these promoters are free of nucleosomes. Another histone mark H2K9ac was shown to be associated with low nucleosome density in human core promoters (Nishida et al. 2006). The ground state of eukaryotic transcription is generally restrictive, such that chromatin-modifying activities must precede the binding of transcription factors to promoters and long-range elements (Struhl 1999). Because chromatin in bidirectional promoters is more open than that for unidirectional promoters, an attractive hypothesis is that bidirectional genes are less restrictive than genes of unidirectional promoters and are either broadly transcribed or poised for transcription, already open and requiring the binding of only a single or a few TFs for transcription of each pair of divergent genes. Consistent with this argument, ChIP-chip data with two chromatin-remodeling complexes, BRG1 associated factor 155 and 170 (SMARCC1 and SMARCC2), indicate that they are underrepresented in bidirectional promoters.

Our finding that GABPA binds to at least 57% of bidirectional promoters in the human genome is strong evidence that it is a major regulator of bidirectional transcription. It is likely that GABPA binds to an even larger fraction of bidirectional promoters, as some GABPA binding sites can be suppressed by methylation and therefore missed in the cell types we studied here. Because it has been shown that GABPA does not bind to its recognition site when methylated (Yokomori et al. 1995, 1998), and because the activities of CpG-rich promoters—a class to which almost all bidirectional promoters belong—are regulated by methylation (Reik et al. 2001; Fazzari and Greally 2004), epigenetic mechanisms almost surely must play a significant role in bidirectional transcription. Methylation status may explain why truncating GABPA site-containing promoter regions did not lead to a severe decrease in bidirectional promoter activities in five of the nine cases that we tested by deletion analysis.

The most obvious explanation for the abundance of bidirectional promoters in mammalian genomes may be that this is a simple and elegant arrangement of gene pairs that need to be coregulated. Indeed, divergent genes have more correlated expression profiles than randomly paired genes. To maximize the

extent of coregulation, one could envision TF motifs exclusive to subsets of bidirectional promoters, which could be activated in distinct cell types under specialized conditions. In contrast, our study reveals that bidirectional promoters have not evolved to optimize the fine control of individual gene pairs. Rather, as a group, bidirectional promoters are likely to be regulated by a limited set of TFs and divergent genes have positively correlated expression patterns. We suggest that this is a simple mechanism for regulating genes that are generally and broadly transcribed.

Methods

Two bidirectional promoter data sets

The first set of bidirectional promoters was selected to maximize quality and used for the ab initio motif search. TSS definitions were taken from DBTSS version 4.0 (Yamashita et al. 2006), and the 500-bp or shorter intergenic regions between TSSs of oppositely stranded nonoverlapping transcripts were taken to be bidirectional promoters. This results in 376 bidirectional promoters.

The second set of bidirectional promoters was selected to increase quantity and used for binding-site enumeration and motif categorization. The 1-kb or shorter intergenic regions between oppositely stranded nonoverlapping transcription units (TUs) were taken to be bidirectional promoters. TUs were constructed by combining TSS annotations from three databases in order of preference: (1) DBTSS (Yamashita et al. 2006), (2) hg16 RefGene table from UCSC (Karolchik et al. 2003), and (3) PromoSer (Halees and Weng 2004). For those transcripts that were neither in DBTSS nor RefSeq, we used the 5′-most PromoSer TSS mappings with quality >2, which correspond to TSSs based on GenBank mRNAs excluding ESTs. Each transcript was then consolidated into TUs defined by PromoSer cluster identifiers. A total of 32,349 TUs were identified, and of the following composition: 9767 derived from DBTSS, 5228 derived from RefGene, and 17,354 derived from PromoSer. This results in 1304 bidirectional promoters. We believe that this larger bidirectional promoter set balances quality and quantity in the selection process.

One unidirectional promoter data set

Using the previously described transcription units, we annotated 13,205 unidirectional promoters whose closest TU neighbor is >1 kb away. A total of 6503 unidirectional promoters are oriented in the forward genomic direction, while 6702 are oriented in the reverse genomic direction.

Gene ontology analysis

The names of divergent gene pairs from the second set of bidirectional promoters were extracted and analyzed using GO::TermFinder (Boyle et al. 2004). GO::TermFinder analyzes a list of genes to determine whether any GO terms occur more frequently than would be expected by chance. The *P*-value was calculated using the hypergeometric distribution and the Bonferroni correction was applied to correct for multiple testing. We set a corrected *P*-value cutoff of 0.1 and obtained a limited list of enriched GO terms available in Supplement 5.

Motif discovery

Sequences from the first bidirectional promoter data set were extracted and analyzed using MEME Version 3.5.0 (Grundy et al. 1996). We requested output of 15 motifs and further narrowed this list down to seven unique motifs by pairwise correlation of the 15 motifs (Haverty et al. 2004). Two motifs were poly(A) and poly(T) repeats, and six motifs were reverse complements of other motifs in the set.

Motif representation calculation

Along with the ab initio discovered motifs, a set of known vertebrate PSSMs were extracted from TRANSFAC (Wingender et al. 1996; Fu and Weng 2005), and mapped onto the large bidirectional promoter set using POSSUM (http://zlab.bu.edu/~mfrith/possum/). POSSUM calculates the log likelihood ratio of observing a subsequence given the motif definition versus observing the subsequence given the nucleotide composition in a 100-bp window around the subsequence. For background comparison, we individually checked the CpG dinucleotide content of every bidirectional promoter and sampled without replacement for a corresponding unidirectional promoter of the top 10 unidirectional promoters with closest CpG dinucleotide content. This results in an average CpG dinucleotide content of 7.42% in our random samples compared with 7.28% in all bidirectional promoters, and a C+G content of 61.83% in our random samples compared with 61.54% in all bidirectional promoters. The binding sites for all of the motifs were enumerated at different POSSUM log likelihood score cutoffs in all bidirectional promoters as well as the random samples. We averaged the classifications obtained by two methods to place motifs in the overrepresented, shared, or underrepresented category. The first method by voting determines whether the motif counts in bidirectional promoters (dotted line in Fig. 2) are predominantly (1) above the unidirectional sum (solid black line), (2) between the unidirectional sum and the single unidirectional counts (solid black and solid gray line, respectively), or (3) below the single unidirectional counts (solid gray line). The second method by distance calculation determines whether the motif counts in bidirectional promoters are closer to (1) the unidirectional sum, (2) the single unidirectional counts, or (3) the horizontal line with a height of 0, by summing the Euclidian distances between the motif counts at each log likelihood cutoff. Overrepresented, shared, and underrepresented motifs were coded as 1, 2, and 3, respectively, and the codes from the two methods were averaged across 10 randomized picks of unidirectional promoter sets. Overrepresented motifs have an average code ≥ 2.5, shared motifs have an average code <2.5 and ≥ 2, and underrepresented motifs have an average code <2. The detailed categorizations and statistical significance of overrepresentation are provided in Supplement 2.

ENCODE ChIP-chip factors in bidirectional promoters

Genomic target lists for various ChIP-chip experiments were reported at 10% false discovery rate cutoff by the ENCODE Consortium. We provide summary statistics of the target lists we used in Supplement 6. On average, each target was 765 bp in length, although the number of targets varied dramatically depending on the factor. Because the scope of the ChIP-chip experiments was within the ENCODE regions (1% of human genome), we extracted bidirectional and unidirectional promoters in these regions using previously defined TUs that overlap with the ENCODE regions, resulting in 23 bidirectional promoters (46 divergent genes), and 227 unidirectional promoters (provided in Supplement 6). A target is said to overlap a promoter (bidirectional or unidirectional) if at least one base pair of the target overlaps within a 500-bp window around the promoter; hence, a portion of the downstream region is also captured in the overlap analysis. This is important, as histones are depleted at the promoter but not downstream of the TSS. We report the percentage of bidirectional promoters (and unidirectional promoters) that overlap with the target list at every rank cutoff. Factors that are

overrepresented show greater overlap in bidirectional promoters than twice the overlap in unidirectional promoters across 70% of the rank cutoffs (the solid line labeled A in Fig. 4 is above the dash-dot line labeled B 70% of the time).

Histone modification analysis

We analyzed raw ChIP-chip enrichment scores of four modified histones (H3ac, H4ac, H3K4me2, and HeK4me3), all in HeLa cells, after 30 min of retinoic acid stimulation. Our results do not change if we separately analyze or include ChIP signal intensities of modified histones before stimulation. To categorize histone-modification patterns, we averaged the overall signal intensities in the diminished intergenic region and the elevated downstream 1-kb region, and checked whether the average signals of three regions in individual bidirectional promoters (1 kb downstream of the reverse gene, bidirectional promoter region, and 1 kb downstream of forward gene) were closer to the overall elevated or diminished averages. For example, the average H3ac signal intensity is 1.90, 1 kb downstream of the reverse strand gene NM_024298, and 1.29, 1 kb downstream of the forward strand gene NM_024075. The bidirectional promoter regulating both genes has an average signal intensity of 0.80. Comparing each number from these three regions to the elevated (1.54), and diminished (0.91) H3ac signals of all divergent genes, we categorize the first and second region as elevated, and the third region as diminished, because the first and second numbers are closer to 1.54, while the third number is closer to 0.91. The ChIP intensities in the overall diminished intergenic region are typically half that of the overall elevated downstream 1-kb regions (see Supplemental Table S1).

Microarray analysis

We used binary (MAS5 absent/present) expression profiles from Affymetrix U133 array to analyze the overall expression levels of all genes and coexpression levels of divergent genes in the ENCODE regions (Koch et al. 2007). To correlate gene-expression profiles, we used the GNF data set of 156 Affymetrix U133 microarray experiments across 78 human cell types (Su et al. 2004). Because a gene can map to several probesets on the microarray, we computed the mean of all pairwise correlation coefficients between probeset pairs. We did not use the signed absolute maximum as in an earlier study (Li et al. 2006), because this systematically exaggerated the correlations and resulted in bimodal distributions for all gene sets we examined—even the randomly paired genes of unidirectional promoters. The earlier study used randomly paired probesets as the background, and hence, did not reveal this artifact (Li et al. 2006).

Deletion constructs

Binding sites of the ab initio discovered motifs were mapped to eight deletion construct experiments using the *cis*-element prediction program POSSUM (score cutoff = 7), and visualization program MotifViz (Fu et al. 2004). Deletion construct functional assays were previously conducted and reported by Trinklein et al. (2004).

Mutation analysis of 30-bp fragment

We generated 3-bp substitution mutations throughout a 30-bp region in one bidirectional promoter and tested each for promoter activity in both directions. We produced these mutations by using random transversions (Pu → Py; Py → Pu) three bases at a time, and shifting by two bases. Both strands of each mutant construct were synthesized by Operon (standard phosphoramadite synthesis), annealed, cloned, and sequence verified. The transcriptional activity was assayed in HT1080 cells in two directions using the luciferase reporter vector described in Trinklein et al. (2004).

GABPA ChIP-qPCR

Genomic GABPA-binding sites were computationally predicted using POSSUM, and 118 sites stratified by region (bidirectional promoter, unidirectional promoter, or nonpromoter), and binding site score (high, medium, low, or no motif) were randomly selected for testing. Binding sites are called high scoring if they have a POSSUM log likelihood score >9; medium scoring if they have a score between 8 and 9; and low scoring if they have a score between 3 and 8.

Sonicated chromatin from either 2 × 107 K562 cells or 4 × 107 Jurkat cells was incubated for 24 h with 5 µg of mouse monoclonal GABPA antibody (catalog # sc-28312, Santa Cruz Biotechnology) coupled to sheep anti-mouse IgG magnetic beads (Dynal Biotech). The magnetic beads were washed five times with buffer containing 100 mM Tris, 500 mM LiCl, 1% NP-40, and 1% deoxycholate, and once with TE buffer (10 mM Tris at pH 8.0, 1 mM EDTA). The DNA was then eluted by incubating in buffer containing 1% SDS and 0.1 M NaHCO3 at 65° for 1–2 h. After removing the magnetic beads, the eluent was further incubated for 16 h at 65° to reverse the cross-links. A phenol chloroform extraction was performed and the aqueous phase desalted and concentrated using the QIAquick PCR purification kit (QIAGEN).

We performed real-time PCR to measure the enrichment at each promoter by incorporation of SYBR-green. We designed primers around the highest-scoring GABPA motif in a promoter, so that the final amplicon was 60–100 bp in length. Reactions were performed according to Bio-Rad recommendations for detection on the iCycler instrument. For each amplicon, we constructed a standard curve of threshold cycles from 50 ng, 5 ng, 500 pg, and 50 pg of genomic DNA (Roche). We then fit the threshold cycle of ChIP-enriched DNA to determine the quantity of starting template. To determine the fold enrichment for any particular fragment, the quantity of starting template was divided by the average starting quantities of three negative controls.

Acknowledgments

We thank the ENCODE Consortium for the ChIP-chip data sets. In particular, ChIP-chip on all modified histones were from the Bing Ren laboratory at UCSD and ChIP-chip on STAT1, and the SMARCC factors were from the Mike Snyder laboratory at Yale. This work was funded by the ENCODE Consortium grant R01HG03110 from the National Human Genome Research Institute (NHGRI) of the NIH to Z.W. and partly funded by the Training Program in Quantitative Biology and Physiology T32GM008764-06 from the NIH to J.M.L. This work was also funded by the ENCODE Consortium grant U01HG003162 from the NHGRI to R.M.M. P.J.C. was supported by the Stanford Genome Training Program (NHGRI grant T32HG00044).

References

Adachi, N. and Lieber, M.R. 2002. Bidirectional gene organization: A common architectural feature of the human genome. *Cell* **109:** 807–809.

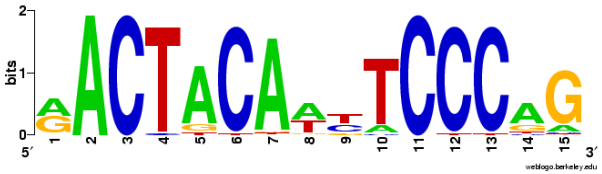
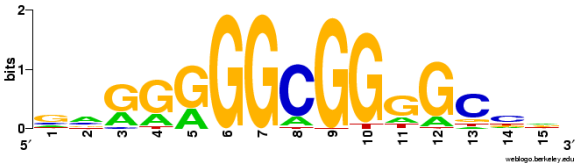
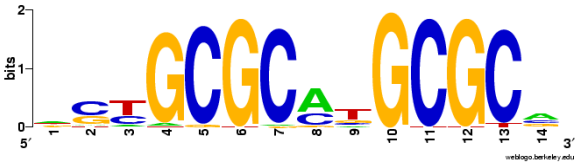
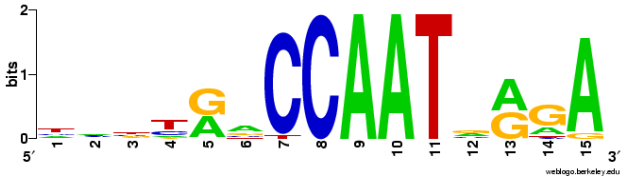
Allfrey, V.G., Faulkner, R., and Mirsky, A.E. 1964. Acetylation and methylation of histones and their possible role in the regulation of Rna synthesis. *Proc. Natl. Acad. Sci.* **51:** 786–794.

Bernstein, B.E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D.K., Huebert, D.J., McMahon, S., Karlsson, E.K., Kulbokas 3rd, E.J., Gingeras, T.R., et al. 2005. Genomic maps and comparative analysis

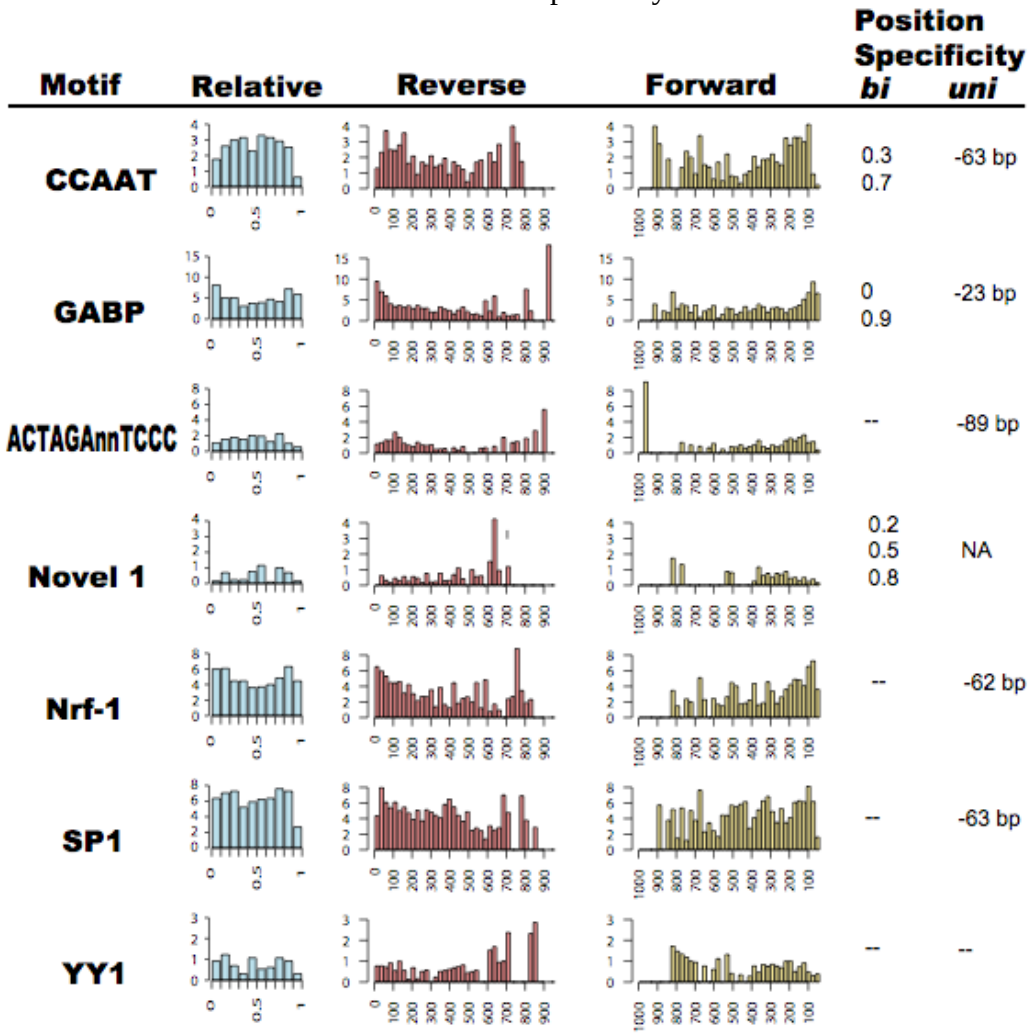
of histone modifications in human and mouse. *Cell* **120:** 169–181.
Bieda, M., Xu, X., Singer, M.A., Green, R., and Farnham, P.J. 2006. Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. *Genome Res.* **16:** 595–605.
Boyle, E.I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J.M., and Sherlock, G. 2004. GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* **20:** 3710–3715.
Bush, T.S., St. Coeur, M., Resendes, K.K., and Rosmarin, A.G. 2003. GA-binding protein (GABP) and Sp1 are required, along with retinoid receptors, to mediate retinoic acid responsiveness of CD18 (β 2 leukocyte integrin): A novel mechanism of transcriptional regulation in myeloid cells. *Blood* **101:** 311–317.
Carninci, P. 2006. Tagging mammalian transcription complexity. *Trends Genet.* **22:** 501–510.
Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., et al. 2005. The transcriptional landscape of the mammalian genome. *Science* **309:** 1559–1563.
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.
Carter, R.S. and Avadhani, N.G. 1994. Cooperative binding of GA-binding protein transcription factors to duplicated transcription initiation region repeats of the cytochrome c oxidase subunit IV gene. *J. Biol. Chem.* **269:** 4381–4387.
The ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306:** 636–640.
The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* (in press).
Engström, P.G., Suzuki, H., Ninomiya, N., Akalin, A., Sessa, L., Lavorgna, G., Brozzi, A., Luzi, L., Tan, S.L., Yang, L., et al. 2006. Complex Loci in human and mouse genomes. *PLoS Genet.* **2:** e47.
Ettwiller, L., Paten, B., Souren, M., Loosli, F., Wittbrodt, J., and Birney, E. 2005. The discovery, positioning and verification of a set of transcription-associated motifs in vertebrates. *Genome Biol.* **6:** R104.
Fazzari, M.J. and Greally, J.M. 2004. Epigenomics: Beyond CpG islands. *Nat. Rev. Genet.* **5:** 446–455.
Fu, Y. and Weng, Z. 2005. Improvement of TRANSFAC matrices using multiple local alignment of transcription factor binding site sequences. *Genome Inform. Ser. Workshop Genome Inform.* **16:** 68–72.
Fu, Y., Frith, M.C., Haverty, P.M., and Weng, Z. 2004. MotifViz: An analysis and visualization tool for motif discovery. *Nucleic Acids Res.* **32:** W420–W423.
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.* (this issue) doi: 10.1101/gr.5533506.
Gopalakrishnan, L. and Scarpulla, R.C. 1995. Structure, expression, and chromosomal assignment of the human gene encoding nuclear respiratory factor 1. *J. Biol. Chem.* **270:** 18019–18025.
Grant, P.A. 2001. A tale of histone modifications. *Genome Biol.* **2:** Reviews0003.1–0003.6.
Grundy, W.N., Bailey, T.L., and Elkan, C.P. 1996. ParaMEME: A parallel implementation and a web interface for a DNA and protein motif discovery tool. *Comput. Appl. Biosci.* **12:** 303–310.
Halees, A.S. and Weng, Z. 2004. PromoSer: Improvements to the algorithm, visualization and accessibility. *Nucleic Acids Res.* **32:** W191–W194.
Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., et al. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32:** D258–D261.
Haverty, P.M., Hansen, U., and Weng, Z. 2004. Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification. *Nucleic Acids Res.* **32:** 179–188.
Kabat, J.L., Barberan-Soler, S., McKenna, P., Clawson, H., Farrer, T., Zahler, A.M., and Stormo, G. 2006. Intronic alternative splicing regulators identified by comparative genomics in nematodes. *PLoS Comput. Biol.* **2:** e86.
Kapranov, P., Drenkow, J., Cheng, J., Long, J., Helt, G., Dike, S., and Gingeras, T.R. 2005. Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res.* **15:** 987–997.
Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., et al. 2003. The UCSC Genome Browser Database. *Nucleic Acids Res.* **31:** 51–54.
Koch, C.M., Andrews, R.M., Flicek, P., Karaoz, D.S.C.U., Clelland, G.K., Wilcox, S., Beare, D.M., Fowler, J.C., Couttet, P., et al. 2007. The Landscape of Activating Histone Modifications Across 1% of the Human Genome. *Genome Res.* (this issue) doi: 10.1101/gr.5704207.
Koyanagi, K.O., Hagiwara, M., Itoh, T., Gojobori, T., and Imanishi, T. 2005. Comparative genomics of bidirectional gene pairs and its implications for the evolution of a transcriptional regulation system. *Gene* **353:** 169–176.
Li, Y.Y., Yu, H., Guo, Z.M., Guo, T.Q., Tu, K., and Li, Y.X. 2006. Systematic analysis of head-to-head gene organization: Evolutionary conservation and potential biological relevance. *PLoS Comput. Biol.* **2:** e74.
Mantovani, R. 1999. The molecular biology of the CCAAT-binding factor NFY. *Gene* **239:** 15–27.
Nishida, H., Suzuki, T., Kondo, S., Miura, H., Fujimura, Y., and Hayashizaki, Y. 2006. Histone H3 acetylated at lysine 9 in promoter is associated with low nucleosome density in the vicinity of transcription start site in human cell. *Chromosome Res.* **14:** 203–211.
Oei, S.L. and Shi, Y. 2001. Transcription factor Yin Yang 1 stimulates poly(ADP-ribosyl)ation and DNA repair. *Biochem. Biophys. Res. Commun.* **284:** 450–454.
Patton, J., Block, S., Coombs, C., and Martin, M.E. 2005. Identification of functional elements in the murine Gabpα/ATP synthase coupling factor 6 bi-directional promoter. *Gene* **369:** 35–44.
Reik, W., Dean, W., and Walter, J. 2001. Epigenetic reprogramming in mammalian development. *Science* **293:** 1089–1093.
Scarpulla, R.C. 2006. Nuclear control of respiratory gene expression in mammalian cells. *J. Cell. Biochem.* **97:** 673–683.
Struhl, K. 1999. Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell* **98:** 1–4.
Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci.* **101:** 6062–6067.
Takai, D. and Jones, P.A. 2004. Origins of bidirectional promoters: Computational analyses of intergenic distance in the human genome. *Mol. Biol. Evol.* **21:** 463–467.
Trinklein, N.D., Aldred, S.F., Hartman, S.J., Schroeder, D.I., Otillar, R.P., and Myers, R.M. 2004. An abundance of bidirectional promoters in the human genome. *Genome Res.* **14:** 62–66.
Wingender, E., Dietze, P., Karas, H., and Knuppel, R. 1996. TRANSFAC: A database on transcription factors and their DNA binding sites. *Nucleic Acids Res.* **24:** 238–241.
Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., and Kellis, M. 2005. Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **434:** 338–345.
Yamashita, R., Suzuki, Y., Wakaguri, H., Tsuritani, K., Nakai, K., and Sugano, S. 2006. DBTSS: Database of human transcription start sites, progress report 2006. *Nucleic Acids Res.* **34:** D86–D89.
Yokomori, N., Kobayashi, R., Moore, R., Sueyoshi, T., and Negishi, M. 1995. A DNA methylation site in the male-specific P450 (Cyp 2d-9) promoter and binding of the heteromeric transcription factor GABP. *Mol. Cell. Biol.* **15:** 5355–5362.
Yokomori, N., Tawata, M., Saito, T., Shimura, H., and Onaya, T. 1998. Regulation of the rat thyrotropin receptor gene by the methylation-sensitive transcription factor GA-binding protein. *Mol. Endocrinol.* **12:** 1241–1249.

Received June 12, 2006; accepted in revised form January 2, 2007.

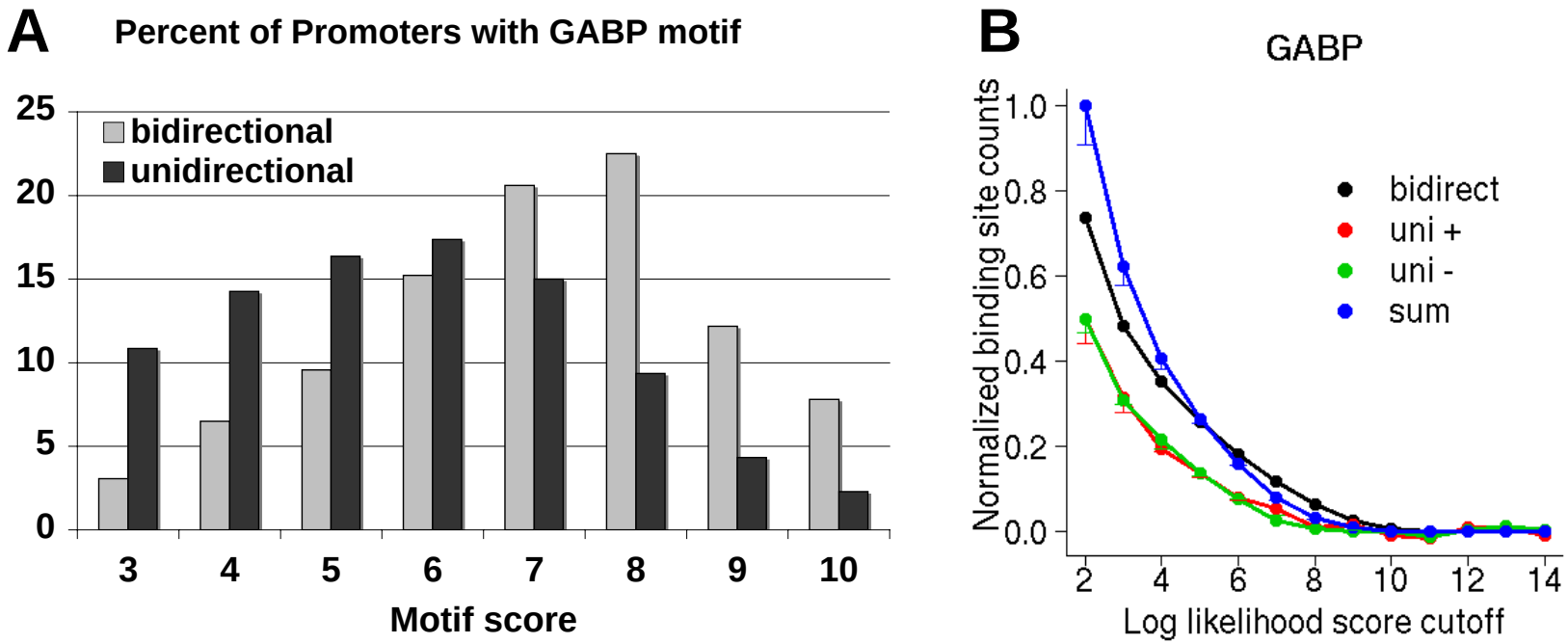
Supplement 1, Figure S1. Sequence Logos of unique motifs discovered by *ab initio* method, and corresponding TRANSFAC PSSMs.

	***Ab initio* Motif**	**Corresponding TRANSFAC Motif**
ACTACAnnTCCC		
GABP		
SP1		
YY1		
Nrf-1		
CCAAT		
Novel 1		

Supplement 1, Figure S2. Position specificity histograms of *ab initio* motifs in bidirectional promoters normalized by the number of promoters that span each bin. For all the plots, the y-axis represents the number of binding sites with log likelihood score greater than 7, and the x-axis represents the span of the bidirectional promoter either in "relative" units or absolute base pairs. The second column, labeled "Relative," shows the positions relative to both flanking TSSs. The third column, labeled "Reverse," shows the position in bp relative to the TSS of the minus strand gene, and the fourth column labeled "Forward" is relative to the TSS of the plus strand gene. Position specificity (in relative position and bp) of bidirectional promoters and unidirectional promoters are summarized in the fifth and sixth column respectively.

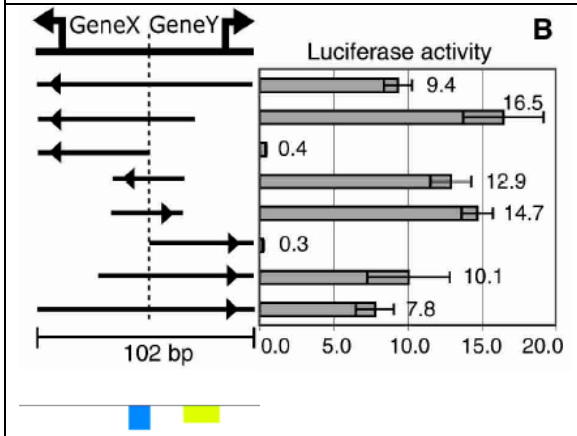


Supplement 1, Figure S3. GABP binding site presence by proportion (A) and by raw frequency (B). (A) Percent of bidirectional (1,304) and unidirectional (13,205) promoters containing at least one GABP binding site at various motif score cutoffs. Recall high scoring sites have score 9 or higher, medium scoring sites are between 8 and 9, while low scoring sites are between 3 and 8. A larger proportion of bidirectional promoters contain medium and high scoring sites, whereas a larger proportion of unidirectional promoters contain low scoring sites. **(B) Binding site frequencies normalized by the maximum count in the plot.** Bidirectional promoters contain more GABP binding sites with score greater than 6.



Supplementary figure S4. Deletion Construct Experiments with GABP Binding Site Mappings. The height of the bars are proportional to the log-likelihood score of the motif, which is provided in the rightmost column labeled "score". Bars above the black line are motifs in the plus orientation, while bars below the black line are motifs in the minus orientation.

A. BC021744 AK022122

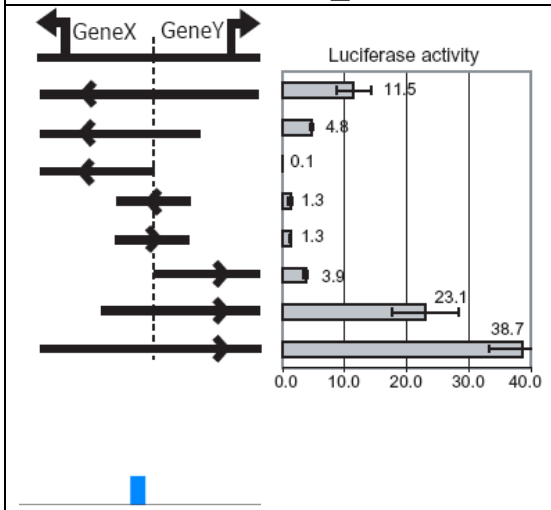


101 bp

sequence	position	strand	motif	score
ccggaaatg	47 - 55	-	GABP	7.92
tctgtttggatgtgg	70 - 84	-	CCAAT	5.61

```
  1 AAAACCGACGGAGGACATAATTTCCAGCCGACCTAAACGAAGGGAACCGG  50
 51 AAATGCGTCATAGCGCGCGTCTGTTTGGATGTGGAAGCCGAGACCTAAAG 100
101 T 101
```

B. BC026014 NM_052951

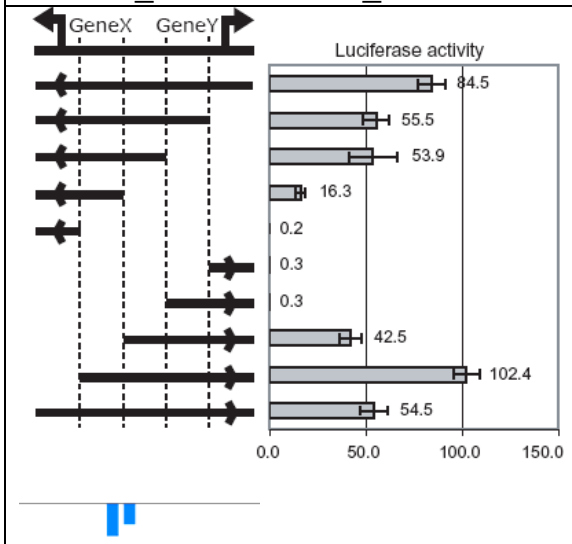


150 bp

sequence	position	Strand	motif	score
cgcttccgg	70 - 78	+	GABP	8.64

```
  1 CTAAGTGTAACTGTCCTGGGCGAGGCGCGGCGGTTCGGTTCCCATGGTAA  50
 51 CCCCGCAGCTCCAGCGTCGCGCTTCCGGGCGGACGAGCAGCGCGCTCCAG 100
101 TGACGTCACGGCGCCACTTTCCGGCCGGTGACAGAGTCCAGCGGAGTTGT 150
```

C. NM_016145 NM_032332

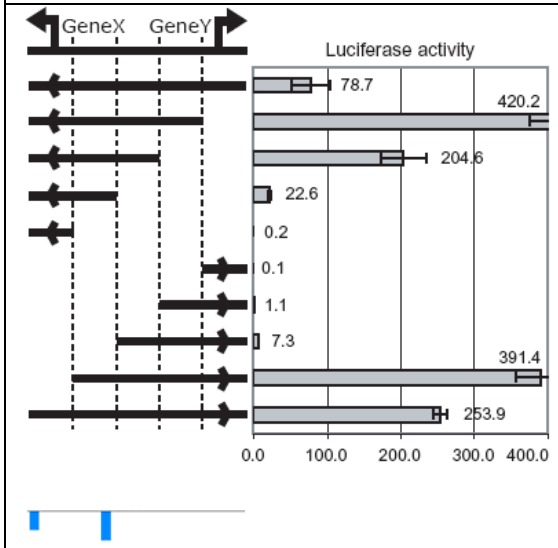


200 bp

sequence	position	strand	motif	score
ccggaagtg	74 - 82	-	GABP	10.6
caggaagta	88 - 96	-	GABP	6.75

```
  1 AAAATGGGGAAAAGAACCTGAAAGCGGGCTTCAATAGTTCCAACCCGATC 50
 51 ACCGAAATGCCAAAATGATGAACCCGGAAGTGATAAACAGGAAGTAGGTC 100
101 AGGAAGAACAATAATAAGGTTACACCCAAGCGTGGGTTTCTAAGGCGCGG 150
151 AATTTTCCGTACAGACCGATTTAAGGCTGCAAGGAAGGAGTCCTGGGAGC 200
```

D. AK024664 NM_005927



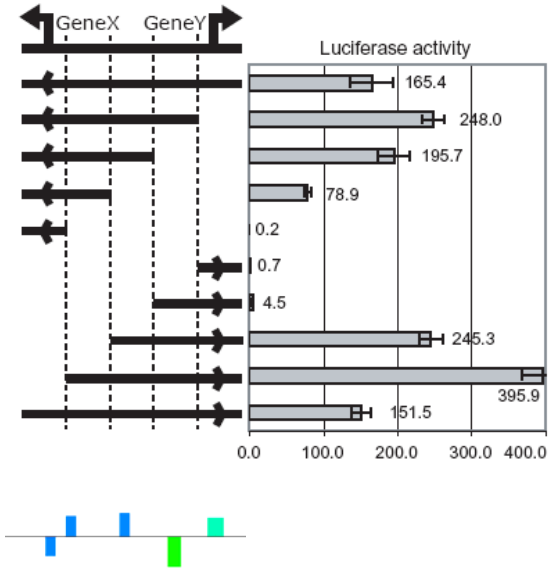
219 bp

sequence	position	strand	motif	score
caggaagct	4 - 12	-	GABP	6.41
ccggaagtg	76 - 84	-	GABP	10.0

```
  1 GTTCAGGAAGCTCCGCACCTATCCGGCCGCCACCCTCAGCACAGCCACGG 50
 51 CAGCCGACTCGGCGTTCCTACTGCCCCGGAAGTGCTCCTTCAGCGCAGAG 100
101 GCGTGCCGGTGTGTCCGAGAAAACTTCCGAGTTAAGCCGCCGCTGAGGCC 150
151 GGAAGGAGCTAGACGGCGGTCGGGTAGGTGACGGCTTGGTTGGGGTCGCT 200
201 GCGAGGGGACTACTAGGAG 219
```

E. AK027016 NM_014039

225 bp

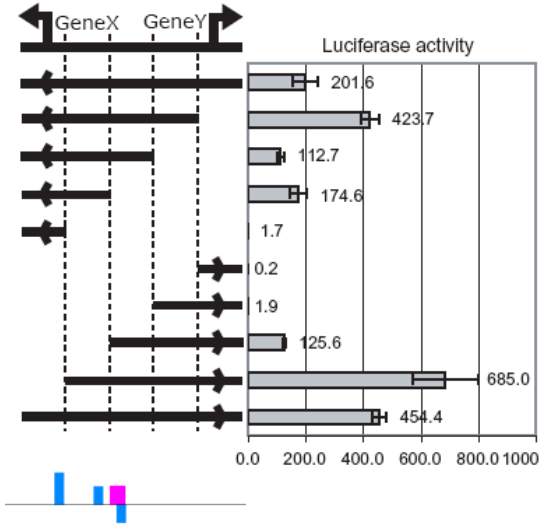


sequence	position	strand	motif	score
gcggaagaa	39 - 47	-	GABP	6.22
tatttccgt	58 - 66	+	GABP	6.38
agcttccgg	108 - 116	+	GABP	7.31
ggcgccatttttt	153 - 164	-	YY1	9.64
acgtgggcggagctc	190 - 204	+	SP1	5.83

```
  1 CTCCAACTCCCGATAACCAGCCGACCTCCTCCAACCGTGCGGAAGAAAAG 50
 51 GGTTGGCTATTTCCGTGGCCCAAGTAATAGTCAGGCCGAAATCTCGCGAT 100
101 ACAACTTAGCTTCCGCGAAGAAGGCGAGCGGTGGGAGGAGACGCGACGTG 150
151 GGGGGCGCCATTTTTCTCGCCGCGCAGGGAGGACTGACTAACGTGGGCGGA 200
201 GCTCTAGCTCGCGTATTCTGAGGAG 225
```

F. AK025314 BC005288

239 bp

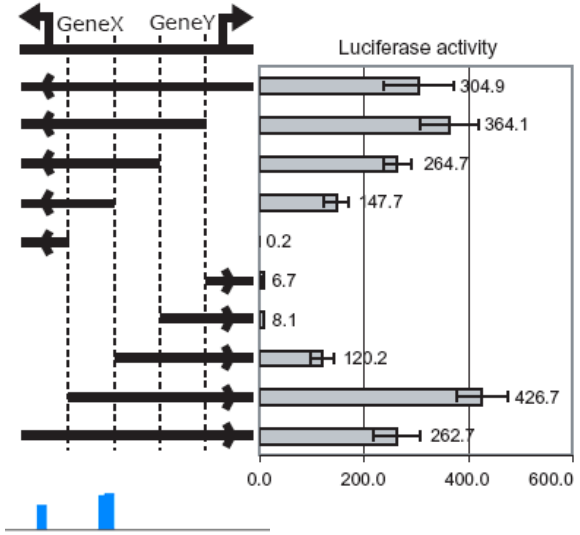


sequence	position	strand	motif	score
cacttccgg	50 - 58	+	GABP	10.0
aacttccgt	89 - 97	+	GABP	5.58
gctaaaagcggaaaa	105 - 119	+	Novel_1	5.89
gcggaaaac	112 - 120	-	GABP	5.80

```
  1 AAGAAAATAGGCTTTCTCCGCTCTACCGCCTCGGGCAGCCACACCTCCAC 50
 51 ACTTCCGGCGGTGTACCGGCCAAATGCCGCCTGCCAGCAACTTCCGTCCT 100
101 CCTAGCTAAAAGCGGAAAACAGAGGCTCGGAACCGCTGCGTGGTTCTTGC 150
151 TCTTCACTCGGCCGTTTTAAAGGGTGACTCTTTCCTGTCCCGGCCTGCGT 200
201 GGTGTGGGCTTGTGGGTCTTTGAGACCCGAAAATTGAGA 239
```

G. BC000104 BC009228

255 bp

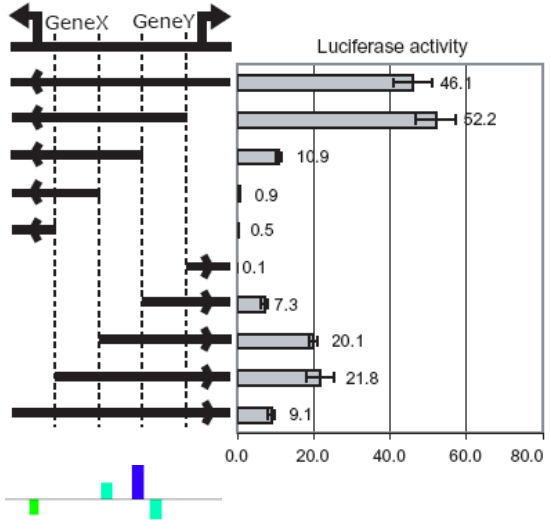


sequence	position	strand	motif	score
ggcttccgt	32 - 40	+	GABP	7.07
cacttccgc	91 - 99	+	GABP	9.65
cgcttccgg	97 - 105	+	GABP	10.2

```
  1 GCTCTACAGCCAGGAAAAGTGAAGCGAAAACGGCTTCCGTAGACTCCGCC  50
 51 ACCACCGAGTAACAGACCAACTCTGACAGCCCGAAGATCGCACTTCCGCT 100
101 TCCGCGGCATAGCACCGGAAAGGCTCGTATCGGAGAACGGTGTGGCCACG 150
151 AGAGTCTAGAGGAAACCAACCCGCAACTCACTGGAGCTGAAGGGCTGGTA 200
201 GTCAGGCCGAATCTCGCGAGAATTGGAACTATTTCAGAGCCAGAGGAGGA 250
251 TAGTG 255
```

H. BC001870 NM_014748

293 bp

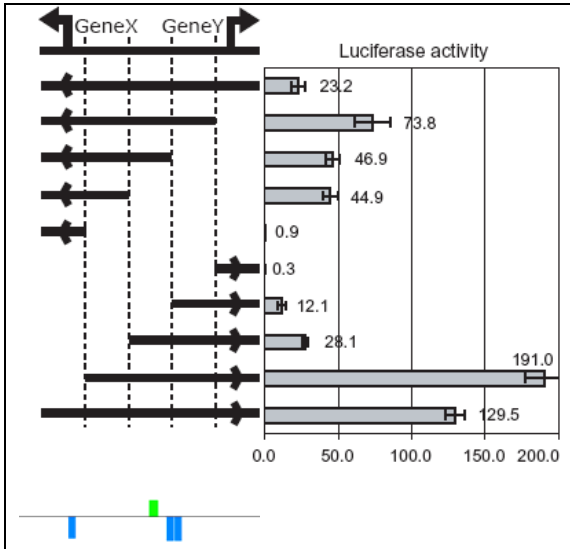


sequence	position	strand	motif	score
gcagccatcgcc	34 - 45	-	YY1	5.15
ctggaggctgagccg	131 - 145	+	SP1	5.43
gactacaatttccga	173 - 187	+	ACTACAnnTCCC	11.5
aggccccgccccgca	197 - 211	-	SP1	6.52

```
  1 CTTCACTCACCCTCGCGAACAGCCACGGCCACAGCAGCCATCGCCCTCAG  50
 51 TCCTAGGCTCCGCCCGCCGTGGTCACGGAGCCGCCCCGCTGCCTCGCTCC 100
101 GCGGCTGCAGGGTCCACAGCGCCACCTGGGCTGGAGGCTGAGCCGCAGAA 150
151 GCGAAGCGCGCAGCCCCAATCGGACTACAATTTCCGACGGGCCCCGAGGC 200
201 CCCGCCCCGCACTACGGGTTGCAATGAGGGCCAACTCCACTTTGGGCTCC 250
251 CTATCCGGACAGGTGGCTCTTGCCCTTTAGACTACAGTTCCCA 293
```

I. NM_020154 AK057200

331 bp



sequence	position	strand	motif	score
ccggaagcc	69 - 77	-	GABP	6.96
gatcatggccga	180 - 191	+	YY1	5.13
gcggaagta	204 - 212	-	GABP	7.70
gcggaagta	215 - 223	-	GABP	7.71

```
  1 AGCAGGACGGGAAAGAAGCCCCACAGAGCGGCCGCCATGACAGCAGCTCT 50
 51 GCACTCAGACCGGCAGCCCCGGAAGCCCTCCCAGTCTCCCGGCCGCCACC 100
101 GCCGCCGCTGGCGGCTTGATGTTGTCATCACTGCGCGCCCGTTTCTGTCA 150
151 TAAGCTGTGGGATCCAACCAGAGGGAAGAGATCATCGGCCCATCGACTAGC 200
201 GAGGCGGAAGTAAGGCGGAAGTACTGCCCTTAACTCCCGGCACTCTGGAT 250
251 TGGAAGACGGCGCACAGAGCCTGTGTCAAGTGCCTATGAAGGGGACTGCC 300
301 CATGAAGTGAAAGTCAAGTGTGTGTTGCTGC 331
```

Supplement 1, Table 1. Average histone modification ChIP intensities in four regions: (1) 1kb downstream of backward gene, (2) bidirectional promoter, (3) 1kb downstream of forward gene, and (4) unidirectional promoter.

modification	avg_left_1kb	avg_bi_prom	avg_right_1kb	avg_uni_prom
UCSDNg_H3K4me2_HeLa_p0	1.266378847	0.548789964	1.36619785	1.0557688
UCSDNg_H3K4me2_HeLa_p30	1.193545456	0.535733682	1.248760963	0.9972586
UCSDNg_H3K4me3_HeLa_p0	1.425071122	0.903546004	1.407148416	0.8415018
UCSDNg_H3K4me3_HeLa_p30	1.178747261	0.707584519	1.274631656	0.7141955
UCSDNg_H3ac_HeLa_p0	1.279876906	0.738174977	1.41320123	1.0151013
UCSDNg_H3ac_HeLa_p30	1.472924124	0.9124908	1.598473821	1.1882837
UCSDNg_H4ac_HeLa_p0	0.499296079	0.34012999	0.607546409	0.5376639