CisView: A Browser and Database of *cis*-regulatory Modules Predicted in the Mouse Genome

Alexei A. Sharov, Dawood B. Dudekula, and Minoru S. H. Ko*

Developmental Genomics and Aging Section, Laboratory of Genetics, National Institute on Aging, National Institutes of Health, 333 Cassell Drive, Suite 3000, Baltimore, MD 21224, USA

(Received 8 May 2006; revised 23 July 2006; published online 15 September 2006)

Abstract

To facilitate the analysis of gene regulatory regions of the mouse genome, we developed a CisView (http://lgsun.grc.nia.nih.gov/cisview), a browser and database of genome-wide potential transcription factor binding sites (TFBSs) that were identified using 134 position-weight matrices and 219 sequence patterns from various sources and were presented with the information about sequence conservation, neighboring genes and their structures, GO annotations, protein domains, DNA repeats and CpG islands. Analysis of the distribution of TFBSs revealed that many TFBSs ($N = 145$) were over-represented near transcription start sites. We also identified potential *cis*-regulatory modules (CRMs) defined as clusters of conserved TFBSs in the entire mouse genome. Out of 739 074 CRMs, 157 442 had a significantly higher regulatory potential score than semi-random sequences generated with a 3rd-order Markov process. The CisView browser provides a user-friendly computer environment for studying transcription regulation on a whole-genome scale and can also be used for interpreting microarray experiments and identifying putative targets of transcription factors.

Key words: transcription factor binding site; evolutionary conservation; promoter; enhancer; CpG island; transcription start site

1. Introduction

The analysis of regulatory regions is a major challenge for contemporary genomics, which requires both experimental and computational studies. Several software tools are available for predicting transcription factor binding sites (TFBSs) in DNA sequences, including MatInspector,[1] TRED[2] and MATCH.[3] Browser rVISTA allows cross-species comparison of TFBSs in the promoters of orthologous genes.[4] A database of mouse and human promoters at UIC (PAGen@UIC) contains information on the location of major TFBSs within 2 kb upstream and 250 bp downstream of a transcription start site (TSS).[5] The ECR browser[6] displays conserved TFBSs in the entire mouse genome, but it uses only a small fraction of known TFBS motifs, and there is no information on non-conserved TFBSs, repeats and CpG islands. This browser does not support queries on specific TFBSs in their promoters. Thus, there is no software that combines genome-wide browsing of TFBSs with sufficient context information (e.g. nucleotide composition, CpG islands, repeat type, evolutionary conservation score and neighboring genes) and tools for analysis of promoters. Furthermore, most tools examine only short segments of promoters, which are not long enough to identify regulatory elements in mammalian genomes. Although the UCSC genome browser[7] can incorporate various kinds of DNA annotations including TFBSs, the information on TFBSs is not available for the mouse genome. Besides, the UCSC genome browser does not offer specialized tools for the analysis of regulatory sequences (e.g. highlighting one or several specific TFBSs and searching for pairs of TFBSs at a certain distance).

Functional TFBSs are usually clustered because of the cooperative nature of transcription factors.[8,9] Clusters of functional TFBSs or *cis*-regulatory modules (CRMs) are classified into proximal CRMs (i.e. promoters) and distal CRMs (DCRMs) that include enhancers, silencers and insulators.[10] In contrast to the proximal CRMs, which have been intensively studied and documented, DCRMs are poorly characterized. The only approach implemented

Communicated by Kenta Nakai

* To whom correspondence should be addressed. Tel. +1-410-558-8359; Fax. +1-410-558-8331, Email: kom@mail.nih.gov

for predicting DCRMs in mouse is the analysis of conserved non-coding sequences.[11] To reduce the number of false positives it is important to use additional information including TFBSs as it has been done for *Drosophila*.[12] A database of predicted DCRMs is also required to have common reference points for describing the regulatory regions of specific genes.

In this paper we present a browser CisView to visualize and query the locations and internal structures of regulatory regions in the mouse genome in combination with a database of predicted TFBSs and CRMs (http://lgsun.grc.nia.nih.gov/cisview).

2. Materials and methods

1.2. *Mapping TFBSs on the mouse genome*

TFBSs were predicted in the entire genome using transcription factor binding models (TFBMs) in the form of position-weight matrices or sequence patterns. Most position-weight matrices are based on the TRANSFAC database (public version 7.0).[13] Because TRANSFAC database has many redundant entries, we manually combined 291 vertebrate position-weight matrices into 115 groups based on overlapping gene sets and/or matrix similarity. Also we trimmed regions with low-information content or with inconsistencies between various matrices describing binding sites of the same transcription factor (TF), as it is documented in our web site (e.g. http://lgsun.grc.nia.nih.gov/geneindex/mm6/TFBS/TF_OCT.html). After trimming one matrix was built for each group. A similar effort to group TFBS patterns into a smaller number of families was reported earlier.[14] The second major source of TFBMs was the set of 174 patterns over-represented in conserved regions of mammalian promoters.[15] Out of these 174 patterns, 69 patterns corresponded to known TFs. We also added TFBMs manually from the literature (see the web site for references). In total, 134 matrices and 219 sequence patterns were used to identify potential TFBSs (below we refer to them simply as TFBSs).

Sequence patterns contained both exact matching symbols (A, T, G and C) and degenerate symbols (e.g. R, Y and N). Because mismatches between DNA sequence and a pattern were not allowed, some long patterns generated too few hits. This problem was handled by treating patterns with total information measure for all symbols $\geq$18 bits ($N = 19$) as matrices and allowing mismatches as described below. The information measure of a symbol = 2, 1, 0.42 and 0 bits if it included 1, 2, 3 and 4 possible nucleotides at the same position, respectively.

Search for binding sites characterized by a position-weight matrix started with a search for a core pattern that is the most informative portion of the binding site. A traditional approach to detect the core is based on the maximum information measured for four consecutive positions in the matrix.[1] However, this method may not perform well for binding sites with a long stretch of the same nucleotide or with two groups of most specific positions separated by a gap (e.g. in palindromes). Thus, we developed a new method to identify the core pattern, which consisted of three or four elements characterized by two most dramatic changes in nucleotide frequency between positions measured by

$$c_j = 0.5 \cdot \left(I_j + I_{j+1}\right) \cdot \sum_i |p_{ij} - p_{ij+1}|, \tag{1}$$

where c_j is the degree of change from position j to position $j + 1$, p_{ij} is the frequency of nucleotide i at position j and I_j is the information measure [i.e. $\sum_i p_{ij} \log_2(p_{ij})$] at position j. Core patterns were allowed to contain degenerate symbols and in this case they included nucleotides that occurred at frequencies >50% of the maximum frequency at that position. Some core patterns had two pairs of nucleotides separated by some distance. For example, the TF_CDP binding site had a core ATNNAT. A full list of core patterns can be downloaded from our website. Our method addressed two problems of core patterns: it can generate cores with gaps and it has a strong preference for non-uniform patterns. The match of the core pattern ensured the proper position of the matrix and reduced the number of false positives.

Each match of the core pattern was then examined if it also matched with the entire matrix using a similarity score. The similarity score is equal to the sum of character heights in a sequence logo[16] divided by the sum of maximum heights at all positions, which is equivalent to the score used in the MatInspector.[1] The minimum similarity threshold was allowed to be 0.8 (i.e. 20% mismatch). However, for abundant TFBSs we used higher similarity thresholds adjusted so that the frequency of matches in CpG-rich and CpG-poor semi-random sequences did not exceed 1 per 500 bp and 2000 bp, respectively. We used different thresholds for CpG-rich and CpG-poor semi-random sequences, because CpG islands are enriched in promoters and can be regulated epigenetically[17]; hence, they are more likely to be enriched in functional TFBSs. Semi-random sequences were generated using third-order Markov models with transition probabilities estimated from CpG-rich and CpG-poor regions in mouse promoters. Following the identification of the similarity threshold for each TFBS, the method was then applied uniformly to the entire genome sequence.

2.2. *Identification of promoter regions*

To identify TSSs, we used the following primary data sources: (i) the mouse genome sequence assembled in March 2005 (mm6)[18]; (ii) DBTSS database ver. 5.2, which was compiled from a large set of full-length cDNAs[19,20]; (iii) NIA Mouse Gene Index, ver. mm6,[21,22] which was compiled from all publicly available mouse cDNA sequences (including full-length cDNA sequences from the RIKEN,[23] Mammalian Gene Collection,[24] KAZUSA[25]), NCBI RefSeq,[26] Ensembl transcripts,[27] and expressed sequence tags (ESTs) from dbEST, including the WashU[28] and NIA cDNAs[29]; and (iv) potential TSSs predicted by the FirstEF software, which uses discriminant functions to identify TSSs and potential donor splice sites based on the frequency distribution of short motifs in the DNA sequence.[30]

TSSs were compiled from several databases in an attempt to cover main and alternative transcripts of protein-coding genes in the mouse genome. Depending on the estimated levels of confidence, we arbitrarily divided TSSs into three groups: high-quality, medium-quality and low-quality. TSSs from DBTSS database were considered high-quality because they were identified using a large set of full-length cDNAs[19,20]. Because the current DBTSS database was based on the earlier version of mouse genome (mm5), we used BLAT[31] to remap TSS coordinates to the mouse genome mm6 (N = 18 503 after remapping). Medium-quality TSSs were identified as matches between two independent data sources, if they were >500 bp away from the high-quality TSSs. The first subset (N = 4712) of medium-quality TSSs was taken from protein-coding transcripts (ORF ≥100 amino acids, or known function) in the NIA Mouse Gene Index, ver. mm6,[21,22] if they matched with FirstEF software predictions within 300 bp.[30] We used 300 bp distance as a threshold for matching criterion, because it corresponds to the false discovery rate (FDR)[32] of ∼1% according to the following estimation. If 52 503 TSSs predicted by FirstEF were randomly distributed in the entire genome (3 Gb), then 387 of them in average would appear within 300 bp of 36 829 TSSs identified by aligning mRNA and EST sequences to the genome. Thus, the FDR is equal to 387/36 829 ≅ 1%. The second subset (N = 4219) of medium-quality TSSs was taken from protein-coding transcripts in the NIA Mouse Gene Index if they started within a CpG island but did not match with FirstEF predictions. CpG islands were detected as regions with a minimum of 8 CpG pairs within 250 bp. This threshold was selected based on the frequency distribution of CpG pairs in promoters (shown in the web site). The third subset (N = 27) of medium-quality TSSs was taken from RefSeq sequences if they matched with FirstEF software predictions. Finally, low-quality TSSs (N = 12 960) were taken from the NIA Mouse Gene Index, if they did not match with other data sources.

Recent experimental data showed that many promoters had a cluster of transcription starts rather than a single TSS.[33,34] However, in the current version of CisView we used only one TSS per promoter as identified by DBTSS, NIA Mouse Gene Index or FirstEF, unless TSSs have opposite orientation or are separated by >500 bp. Although it has also been shown that multiple TSSs exist in the promoters,[35] considering all possible transcription starts within a promoter is not feasible currently, because it would increase the size of the database >10 times and cause considerable delays in performing searches. However, most functions of CisView (e.g. finding binding sites within 1 kb upstream of TSSs) are not critically affected by this treatment.

Tentative promoter boundaries for high-quality and medium-quality TSSs were set to the boundary of a CpG island, if it was present at TSSs; otherwise they were assumed to span from −200 to +100 bp. The promoter boundaries were then adjusted by excluding transposon-related repeats and CDS, followed by merging with potential CRMs (see below). Promoters for low-quality TSSs were considered only if they coincided with a potential CRM.

2.3. Identification of CRMs

A potential CRM (below we refer to it simply as a CRM) was defined as a genomic region containing at least four conserved TFBSs within each 200 bp of its length (TFBSs in transposon-related repeats and CDS were not counted). Evolutionary conservation is a reliable indicator of functionality of TFBSs.[36,37] If a CRM overlapped with a promoter, then it was merged with the promoter; if it overlapped with the 3′-UTR of genes, we considered it a 3′-UTR-associated CRM; and all other CRMs were considered as DCRMs. 3′-UTR-associated CRMs often regulate post-transcriptional processes such as mRNA stability and translation[15]; we thus distinguished them from DCRMs, which are most probably involved in the regulation of transcription. Genome conservation scores and repeat coordinates were downloaded from the UCSC database.[38] Conservation score 0.5 was used as a threshold to consider a TFBS conserved.

Presence of high-quality TFBSs (i.e. with a low mismatch rate) as well as multiple TFBSs of the same kind in a CRM are considered as indicators of its function as a transcriptional regulator.[39] Thus, we estimated the Regulatory Potential of a CRM by a Score (RPS), which was a sum of regulatory scores for individual TFBS and regulatory scores for multiple TFBSs of the same kind. Our method of estimating RPS is different from the one by Elnitski et al.[40] We used only one genome (mouse), evolutionary conservation scores and matches of known TFBS patterns, whereas Elnitski et al. used multiple genomes without considering known TFBS patterns. The probability of accidental occurrence of TFBSs within a CRM of length L was estimated as $P = D(s) \cdot L$ where s is the similarity score of the binding site and $D(s)$ is the density of binding sites with a similarity score $\geq s$ in a semi-random sequence generated using third-order Markov process. Depending on whether a TFBS was located in a CpG-rich or CpG-poor region, we used semi-random sequence generated with transition probabilities estimated from CpG-rich or CpG-poor regions in the mouse genome, respectively. A regulatory score for a TFBS was estimated as $[-\log_{10}(P) - 2]$ if $P < 0.01$, or set to 0 otherwise. The probability of accidental occurrence of multiple binding sites of the same kind, P_{m}, was estimated as the product of probabilities of their individual occurrences, P. A regulatory score for multiple TFBSs was estimated as $[-\log_{10}(P_{\mathrm{m}}) - 2]$ if $P_{\mathrm{m}} < 0.01$ or set to 0 otherwise. The RPS, which is a sum of regulatory scores for individual TFBS and multiple TFBSs, was then estimated for all CRMs in the mouse genome. The probability distribution of RPS within CRMs of each size class (from 50 to 150; from 150 to 250; from 250 to 350; ...; >1950 bp) was then compared with the probability distribution of RPS estimated for semi-random sequences of size 100, 200, ..., 1900, >1900 bp (the last class included sequence sizes from 2000 to 3000 bp) generated using third-order Markov process with transition probabilities from CpG-rich or CpG-poor regions. Probability distributions of RPS were very similar for CpG-rich or CpG-poor semi-random sequences (Supplementary Figure S1 is available at www.dnaresearch.oxfordjournals.org), and, thus, we averaged them and used them for estimating P-values and the FDR of RPS in CRMs in the same size class. After sorting all CRMs by increasing P-values we estimated the false discovery rate for i-th CRM as $\mathrm{FDR}_{\,i} = P_iN/i$, where P_i is the P-value for i-th CRM and N is the total number of CRMs.[32] We considered that a CRM had a significantly higher RPS than in semi-random sequences if FDR was ≤ 0.1.

2.4. *Software and web interface development*

The CisView browser uses cgi scripts (Perl) for generating pictures and web pages. To accelerate data processing we created data files, which included all the information on genes, sequences and TFBSs for each 60 kb region. Query tools allow users to search for specific TFBS patterns or their combinations in promoters or in DCRMs, to search for specific genes based on gene symbols, annotations, gene ontology (GO) terms or protein domains,and to search for promoters with different quality and/or with a TATA box. Any list of promoters produced by the query tool or uploaded by a user can be further analyzed for TFBSs over-represented in the list and GO terms and protein domains associated preferentially with the list. Protein domains and GO annotations were identified within the NIA Mouse Gene Index[21] using InterPro[41] and Gene Ontology database.[42] Over-representation of TFBSs in promoters of genes with specific GO annotations was evaluated statistically using z-scores estimated by the hypergeometric distribution and FDR ≤ 0.1. Three regions of the promoter, i.e. the upstream region of TSSs (-1000 to -200 bp), at TSSs (-200 to 50 bp) and the downstream region of TSSs (0 to 500 bp), were analyzed separately. In each promoter we determined the presence of a single TFBS, multiple TFBSs and pairs of TFBSs separated by 100 bp. For each GO term we sorted TFBS models by increasing P-values and then estimated the FDR as described above. In our website we presented only TFBSs with over-representation ratio >1.5.

An example screen shot of the CisView browser is shown in Fig. 1. Data are displayed using seven scales in the same page: (i) whole chromosome, (ii) 3 Mb, (iii) 300 kb, (iv) 60 kb, (v) 4 kb, (vi) 500 bp and (vi) 80 bp (4 kb and 500 bp scales are skipped in Fig. 1). The location of TFBSs is shown starting from the 60 kb scale. It is possible to highlight specific TFBS patterns, filter them by conservation score or similarity/difference threshold and to display new patterns defined by the user. Information on TFBSs is displayed together with background information on exons and CDS locations, TSSs, repeats and CRMs. The DNA sequence is characterized by the frequency of CG pairs (e.g. CpG islands) as well as some other motifs like CCCC/GGGG, which is known to be associated with regulatory regions.[43]

3. Results

3.1. *Genome-wide mapping of TFBSs*

Mapping of TFBSs and TSSs in the entire mouse genome was described in the Materials and Methods section. Genome-wide searches of 353 TFBMs (134 matrices and 219 sequence patterns) resulted in 99.5 matches per 1 kb of genomic sequence on average. However, semi-random artificial genome sequences generated using third-order Markov process showed 94.2 matches per 1 kb, suggesting that up to 95% of the TFBSs could be false positives. This is an inevitable outcome of the shortness of TFBMs and is the source of a major difficulty for the computational identification of TFBSs in the genome. Our goal was to include all potential TFBSs in the database and, thus, the large proportion of possible false positives among putative TFBSs was anticipated. In contrast, we estimated that the proportion of false negatives was relatively small (37%), because out of 2502 TFBSs included in TRANSFAC database, 1573 were identified properly using our algorithm. This proportion may even be over-estimated, because some of these TFBSs were of poor quality. For example, out of 11 TFBMs for a matrix V\$SP1_01, none matched the SP1 pattern identified by Xie et al.[15] and none was identified with our algorithm.

To identify the subset of TFBSs that is more likely to be functional, we provided tools to filter the TFBSs according to user-defined parameters for evolutionary sequence conservation scores, mismatch scores, associations with TSSs, associations with other TFBSs and associations with genes characterized by a specific GO term. Three examples of using such information and their utilities will be described below. The analysis below will also serve as

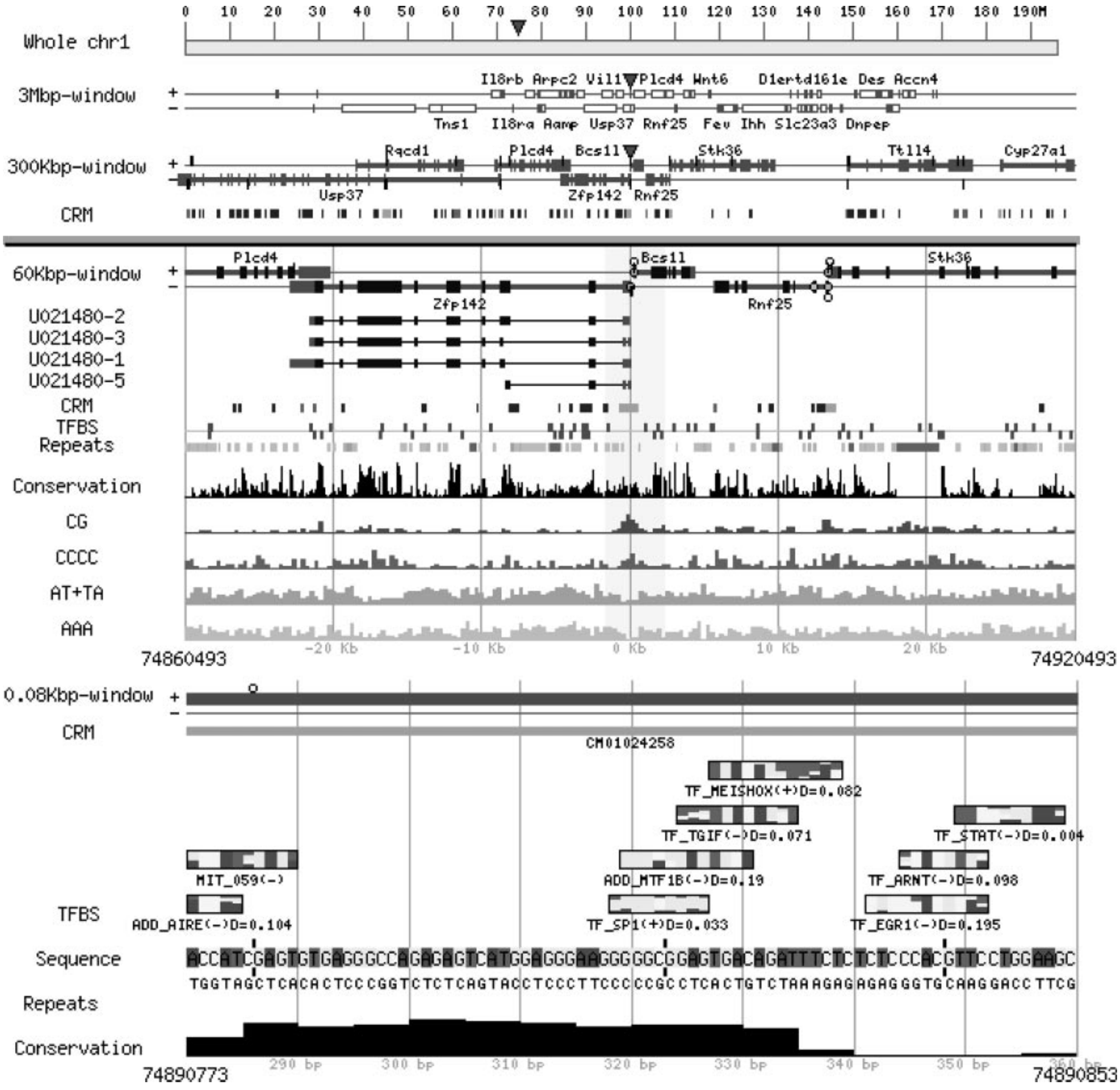

Figure 1. CisView output for genes Zfp142 and Bcs1l (scales for 4 kb and 500 bp are not shown).

the assessment for the quality of TFBSs in the CisView database, although the functionality of TFBSs can only be tested by experimental validations.

3.2. Analyses of TSS-associated TFBSs

The proximity of TFBSs to the TSSs is one of the key parameters for the functionality of TFBSs. We, therefore, analyzed the distribution of TFBSs in the so-called promoter regions that span from −1 kb to +1 kb. Obviously, the analyses will be greatly influenced by the accuracy of TSS locations. In the CisView database/browser, the TSSs were grouped into three categories according to the reliability of data source: 18 503 high-quality TSSs, 8958 medium-quality TSSs and 12 960 low-quality TSSs (see Materials and Methods for the details). Because it is known that TSSs are often associated with an increased density of CpG pairs,[17,44,45] we plotted the distribution of CpG pairs for each TSS group (Fig. 2A). The density of CpG pairs had a clear peak near high-quality and medium-quality TSSs, but no peak near low-quality TSSs. Therefore, we first focused on the high-quality TSSs and identified 122 TFBMs that formed a peak of TFBS abundance in the immediate upstream region of high-quality TSSs (from −100 to 0 bp). The peak of TFBS abundance for these 122 TFBMs was statistically higher (chi square, FDR < 0.05) than the TFBS density in the background (from −1000 to −600 bp and from +600 to +1000 bp) (Supplementary Table S3 is available at www.dnaresearch.oxfordjournals.org). These TFBMs included TF_SP1, ADD_ZF5, MIT_001NRF1, TF_ELK, TF_NFY, TF_AHR, TF_GABP, ADD_WHN, TF_MAZR, TF_ATF1, TF_AP2. Similarly, the density of these TSS-associated TFBSs near the medium-quality TSSs also showed a peak, but as expected there was no peak near low-quality TSSs (Fig. 2B).

To analyze the TSS-associated TFBSs further, we plotted the distributions of TFBSs for individual TFBMs in CpG-rich and CpG-poor promoter regions separately (Fig. 3, plots for all TFBMs are available on our website). Out of the 353 TFBS distributions that we analyzed, 109 showed peaks at a specific distance from the TSSs either in CpG-rich and CpG-poor promoters at least in one orientation (Supplementary Table S4 is available at www.dnaresearch.oxfordjournals.org, Fig. 3). The peaks were significantly higher than the background (FDR $\leq$ 0.1). For example, TF_TATA (TATA-box for Tbp) was considerably more abundant in CpG-poor promoters than in CpG-rich promoters (Fig. 3), which is consistent with the previous report.[46] Some TFBMs showed association with TSSs in CpG-poor promoters (e.g. TF_TATA, MIT_042), whereas others showed association with TSSs in CpG-rich promoters (e.g. TF_SP1, TF_YY1, TF_ETS) (Fig. 3). Most of these peaks were located in a narrow region of the promoter, but some were distributed within the entire CpG islands and showed a broader peak in the plot (Fig. 3, TF_SP1). However, the CpG content alone cannot explain the entire pattern of TF_SP1 distributions. There was a significant over-representation of TF_SP1 sites from −150 to 0 bp, even after removal of the effect of CpG abundance. Interestingly, we noted that 43 TFBSs had a significant orientation bias (FDR $\leq$ 0.10), which exceeded a 2-fold difference for TF_TATA, TF_YY1, MIT_010YY1, TF_CDX, MIT_063, MIT_075STAT, MIT_076AREB6, MIT_095, MIT_102, MIT_0129 and MIT_148.

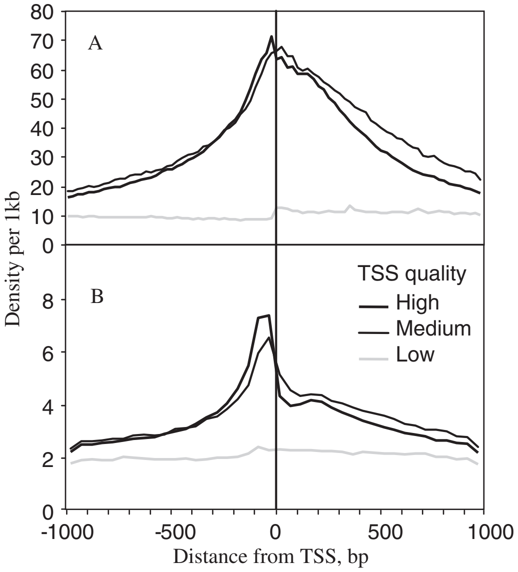

Figure 2. Distribution of CpG pairs (**A**) and TSS-associated TFBSs (**B**) near TSSs.

We combined these two lists of TFBMs (122 TFBMs from Supplementary Table S3 and 109 TFBMs from Supplementary Table S4 are available at www.dnaresearch.oxfordjournals.org), which were derived from different methods of analysis, and obtained 145 TFBMs associated with TSSs. Association with TSSs has been reported earlier for 56 binding patterns by Xie et al.[15] The current study confirmed 51 of these patterns and also identified 72 additional TFBMs (excluding redundancy) for the first time (Supplementary

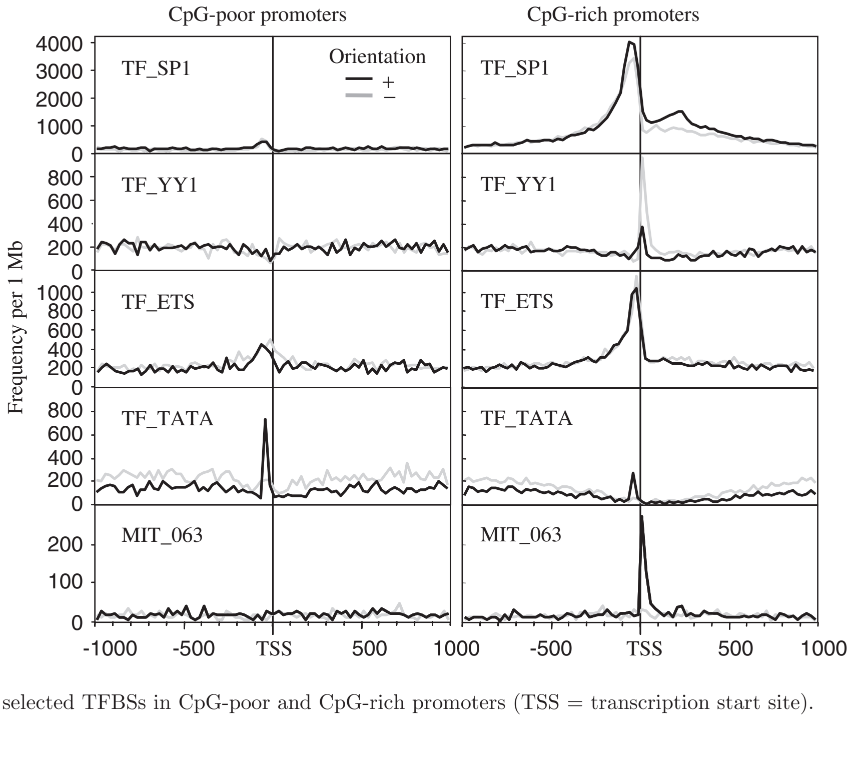

Figure 3. Distribution of selected TFBSs in CpG-poor and CpG-rich promoters (TSS = transcription start site).

Table 1. Summary information on *cis*-regulatory modules (CRMs) in CisView

CRM type	Quality	*N*	Total size (bp)	Average size (bp)	Significant (FDR < 0.1)	Percent significant
DCRM		690 044	166 698 477	242	139 639	20.2
Promoter	High	17 053	10 906 158	640	9350	54.8
Promoter	Medium	7959	4 664 892	586	3769	47.4
Promoter	Low	1599	558 381	349	281	17.6
3′-UTR		22 419	7 574 738	338	4403	19.6
Total		739 074	190 402 646	258	157 442	21.3

Tables S3 and S4 are available at www.dnaresearch.oxfordjournals.org). The analysis and identification of TSS-associated TFBSs for these 145 TFBMs seem to increase one's chance to identify the functional TFBSs located in the promoter regions.

3.3. *Analyses of CRMs*

We assume that predicted TFBSs located within CRMs are more likely to be functional than other TFBSs, because they are evolutionary conserved and clustered. This feature is especially important for TFBSs located far from TSSs, which are known to play an important role in the regulation of mammalian genes, because the association with TSSs cannot be used to filter TFBSs. Therefore, we searched for potential CRMs defined as clusters of conserved TFBSs (see Materials and Methods for the details). Based on the grouping of conserved TFBSs we found 739 074 CRMs, which included 26 611 promoters and 22 419 CRMs associated with 3′-UTR, and 690 044 DCRMs (Table 1). Some promoters ($N = 1991$) included multiple TSSs, the majority of which were bi-directional.[47] There were 17 053 promoters with high-quality TSSs, 7959 promoters with medium-quality TSSs and 1599 promoters with only low-quality TSSs. Coordinates, sequences, regulatory potential scores (RPS) and FDR values for CRMs can be downloaded from the web site (http://lgsun.grc.nia.nih.gov/cisview).

Among all CRMs, 157 442 (21.3%) had a significantly higher (FDR $\leq$ 0.1) RPS than in comparable semi-random sequences. The proportion of significant CRMs was consistent with the quality of promoters assigned in CisView (Table 1). It was highest among the promoters with high-quality TSSs (54.8%) and lowest among the promoters with low-quality TSSs (17.6%). This association also supports our classification of promoter quality.

By definition, DCRMs are not located in the vicinity of TSSs or promoter regions, but they may still be associated with TSSs from a long-range perspective. Thus, we analyzed the association of DCRMs with the TSSs by plotting their frequency distribution in the large region that spans from −30 kb to +30 kb (Fig. 4). A gap in the center of the graph corresponded to the promoter region, which was omitted from this analysis, because it was already analyzed in the Section 3.2. Results showed that the DCRMs had a strong association with TSSs within the distance of 10–15 kb. Interestingly, the DCRMs were more enriched in the downstream regions (e.g. first introns) than in the upstream regions.

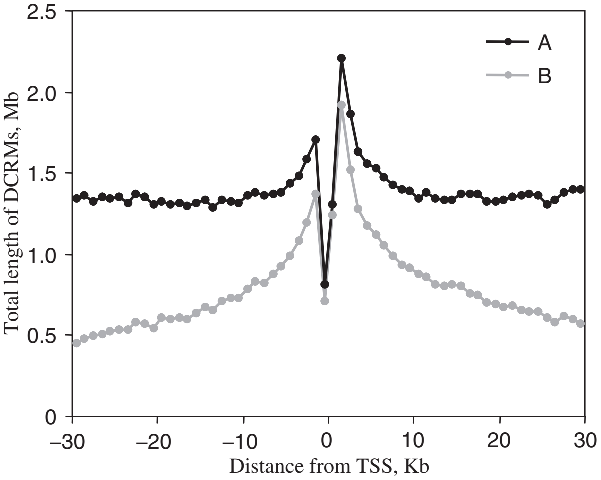

Figure 4. Distribution of the total length of potential distal *cis*-regulatory modules (DCRMs) located at the specific distance from transcription start sites (TSSs) of high and medium quality: (**A**) distance to all TSSs, (**B**) distance to the nearest TSS. The gap in the middle corresponds to promoters which were not counted here.

Out of 39 experimentally tested DCRMs, 32 (82%) were present in the CisView database (Table 2). Seven missing DCRMs had insufficient conservation scores and, thus, were not identified using our algorithm. We, thus, assume that the majority of functional DCRMs are already included in the CisView database, although it may still contain false positives. Only 12 of the validated DCRMs had a significantly higher (FDR $\leq$ 0.1) RPS than in semi-random sequences. This indicates that criteria used for estimating the regulation potential score (presence of rare or high quality TFBSs and multiple TFBSs of the same kind) are not sufficient for a reliable identification of functional CRMs. Thus, we included all predicted DCRMs in the CisView browser, even if their RPS was not significantly higher than in semi-random sequences. Although the RPS is not 100% reliable, we found it useful for comparison of groups of regulatory regions (e.g. promoters of different quality).

Table 2. Known distal *cis*-regulatory modules (DCRMs) and their representation in CisView

Gene	Location (kb)	CRMs in CisView	Reference
Afp	−2.3	CM05009435[a]	59
Afp	−4.8	None	59
Afp	−6.5	None	59
Cdc6	−3	CM11017323	60
Fgf10	−3	CM32000105	61
Foxa2	−15 to −14	CM02022227[a]–29	62
Foxa2	+6 to +11	CM02022216–19[a]	62
Gata4	−38	CM14008199[a]	63
Hoxb2	−1.8	CM11016264[a]	64
Hoxc8	−3	CM15013115–17	65
Lama1	−3.3	None	66
Oct4	−1	CM17005252[a]	53
Oct4	−2	CM17005251	67
Ren1	−2.7	None	68
Sfpi1	−14	CM02029771	69
Ighg	+2	CM12012945	70
Fgf15	−1	CM07022782	71
Myf6	−6	CM10012237	72
Sox2	−4	CM03002384	73
Ccna1	−4.8 to −1.3	CM03031607, CM03004400	74
Phgdh	−1.4	None	75
Cryab	−2.2	CM09006250	76
Pax6	−3.5	CM02015350–51	77
Pax6	−1.5	CM02029880	77
Col1a2	−17 to −15	CM06000113–15, CM06021216	78
Nkx2-5	−5.6	CM17003139[a]	79
Gdf6	−2.3	CM04000545	80
Dnmt3b	−7 to −4	CM02023324	81
Lnp	10	CM02010860	82
Hoxd	−25	CM02010879–80	82
Sry	−5.5	None	83
H19	5	CM07020341	84
H19	6	CM07020339	84
H19	−4 to −2	None	85
Sox9	−28	CM11020867[a]	86
Sox9	−240	CM11045615	86
Otx2	122	CM14023131–32	87
Otx2	−73	CM14023149[a]	87
Myf5	−58 to −56	CM10035736[a]–38	88

[a] RPS is significantly higher (FDR ≤ 0.1) than in semi-random sequences generated using the third-order Markov process.

3.4. Analyses of TFBSs by GO annotations

Gene Ontology (GO) annotations can help to identify the subsets of TFBSs that are most likely functional. For example, TF_E2F binding site was over-represented in promoters of genes involved in the cell cycle (GO:0007049, $N = 117$, http://lgsun.grc.nia.nih.gov/geneindex/mm6/TFBS/go_TF_E2F.html), which is consistent with previous studies.[48] Thus, TF_E2F binding sites that were identified in promoters of these 117 genes are more likely to be functional than those in other promoters.

We also examined TFBSs in promoters of well-established muscle-specific genes.[49] We identified 12 TFBMs over-represented (FDR < 0.05) in 28 promoters of muscle-specific genes[49] from −1000 to +200 bp (Supplementary Table S1 is available at www.dnaresearch.oxfordjournals.org). These TFBMs included TEF, MEF2, MYOD and SRF, which are known to regulate the expression of muscle-specific genes.[50] SP1, which is also considered as muscle-specific transcription factor,[50] had 39 predicted binding sites in these promoters but their over-representation was not statistically significant (FDR = 0.143). Additional TFBMs over-represented in promoters of muscle-specific genes were KLF, AP4, PAX4 and SREBP. Next, we identified 2017 promoters that contained a combination of multiple MYOD and at least one MEF2 binding sites within the region from −1000 to +200 bp. As expected, a top GO term over-represented in the list of corresponding genes was muscle development (GO:0007517, $N = 26$, FDR $<$ 0.001) (Supplementary Table S2 is available at www.dnaresearch.oxfordjournals.org). Other muscle-related GO terms were also significantly over-represented in this list of genes (e.g. contractile fiber, myofibril, sarcomere, actin binding and muscle cell differentiation).

4. Discussion

In this paper we present a CisView, a freely available online browser and database of mouse regulatory regions and TFBSs in the entire mouse genome. The CisView browser provides information not only on TFBS locations but also on genome contexts including neighboring genes, their exon structure, TSSs, CpG-rich regions, repeats of various types, conservation scores, potential CRMs and CpG islands. With easy navigation tools, the browser can be used to find putative gene targets for any transcription factor or a group of transcription factors. Also it can be used to find over-represented TFBSs or pairs of TFBSs in promoters of a given set of genes. The latter feature is useful for the analysis of co-expressed genes identified by microarray analyses. Users can upload their lists of genes or transcripts and analyze them in various ways. Information on GO terms and protein domains for all the genes makes it possible to select a group of genes that belong to a particular pathway (or share the same protein domain) and then to explore common regulatory elements in their promoters or other CRMs. As far as we know, this is the first database and web interface that combine all these analytical tools in a single package.

The CisView database contain TFBMs from the TRANSFAC database (public version 7.0)[13] and various literature sources including reports on individual genes, e.g. Nanog,[51] CTCF[52] and LHR-1.[53] We have implemented a new method to determine the core of the binding motif by assessing the most dramatic change in nucleotide frequencies. The current set of TFBMs covers 85% (78 of 92 TFBMs) of core position-weight matrices in the JASPAR database.[54] In addition to 18 503 high-quality TSSs remapped from DBTSS ver. 5.2, we have identified computationally 8958 medium-quality TSSs, which have very similar distributions of CpG pairs and TSS-associated TFBSs to those of high-quality TSSs. All these medium-quality TSSs are most likely functional, because they are supported by either full-length cDNAs, ESTs or gene models (NCBI and Ensembl). Frequent localizations of many TFBMs (e.g. KLF, NERF, TEL, HTF, ZF5, MAF, CTCF and MAZ) at specific distances from TSSs have been shown for the first time. The database also represents 690 044 DCRMs, which include 82% of known DCRMs based on our estimates. This set of DCRMs can be used for exploring their structure and function both experimentally and computationally. Predicted DCRMs matched well with experimentally determined DCRMs. Also they were over-represented within 10–15 kb from TSSs, which is an indirect indication of their role in the regulation of transcription.

Although we expect that CisView provides a useful platform for the analysis of regulatory regions in the mouse genome, we are also aware of limitations in our approach. A set of motifs that we used for the identification of TFBSs is still incomplete. Currently it accounts for 294 transcription factors. This number may be an underestimation because many paralogous transcription factors use the same binding motif. Nonetheless, it is definitely smaller than the total number of transcription factors ($N = 2068$, NIA Mouse Gene Index, GO:0006355). We also used 105 putative TFBS motifs for which corresponding transcription factors have not been identified yet. Because of the high proportion of false positives, it is important to use additional information on TFBSs, including conservation score, position within a promoter, proximity to other TFBSs, association with functionally related gene sets using tools that are available in the CisView. The method for identification of CRMs, which relies heavily on conservation scores, may not work for non-conserved regulatory regions. To overcome this limitation in the future we plan to use milder criteria of conservation[55] and/or take into account additional information on TFBSs (e.g. similarity scores and association with other TFBSs). The set of analytical tools integrated into CisView can find promoters with a specific combination of up to three TFBSs and identify over-represented TFBSs in a given set of promoters. However, tools for more specialized tasks (e.g. TFBS-based alignment of genomic sequences) are not included. These additional tasks can be handled by other software, e.g. Enhancer Element Locator,[56] Eponine[57] or Genomatix Suite.[58]

In summary, the CisView browser/database provides a user-friendly computer environment for studying transcription regulation on the whole-genome scale. Information on TFBSs is presented with the context of neighboring genes and their structures, GO annotations, protein domains, DNA repeats and CpG islands. Analytical tools include search for genes with a specific combination of TFBSs, identification of TFBSs over-represented in a given set of gene promoters and/or enhancers and plotting the distribution of TFBSs within a set of promoters. CisView can be used for interpreting microarray experiments and identifying putative targets of transcription factors.

Acknowledgements: We thank Ramana V. Davuluri for providing promoter locations of mouse genes generated using their FirstEF program and Michael Zhang for providing the FirstEF software. We would like to thank David Schlessinger, Ramaiah Nagaraja and members of the Ko laboratory for valuable discussion and suggestions. This study utilized the high-performance computational capabilities of the Biowulf/LoBoS3 cluster at the National Institutes of Health. The research was supported by the Intramural Research Program of the National Institute on Aging, NIH.

Supplementary Data: Supplementary data is available online at http://www.dnaresearch.oxfordjournals.org

References

1. Quandt, K., Frech, K., Karas, H., Wingender, E., and Werner, T. 1995, MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data, *Nucleic Acids Res.*, **23**, 4878–4884.
2. Zhao, F., Xuan, Z., Liu, L., and Zhang, M. Q. 2005, TRED: a Transcriptional Regulatory Element Database and a platform for in silico gene regulation studies, *Nucleic Acids Res.*, **33**, D103–D107.
3. Kel, A. E., Gossling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O. V., and Wingender, E. 2003, MATCH: A tool for searching transcription factor binding sites in DNA sequences, *Nucleic Acids Res.*, **31**, 3576–3579.
4. Loots, G. G. and Ovcharenko, I. 2004, rVISTA 2.0: evolutionary analysis of transcription factor binding sites, *Nucleic Acids Res.*, **32**, W217–W221.
5. Kamalakaran, S., Radhakrishnan, S. K., and Beck, W. T. 2005. Identification of estrogen-responsive genes using a genome-wide analysis of promoter elements for transcription factor binding sites, *J. Biol. Chem.*, **280**, 21491–21497.
6. Ovcharenko, I., Nobrega, M. A., Loots, G. G., and Stubbs, L. 2004, ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes, *Nucleic Acids Res.*, **32**, W280–W286.
7. Karolchik, D., Baertsch, R., Diekhans, M., et al. 2003, The UCSC Genome Browser Database, *Nucleic Acids Res.*, **31**, 51–54.

8. Ogata, K., Sato, K., and Tahirov, T. H. 2003, Eukaryotic transcriptional regulatory complexes: cooperativity from near and afar, *Curr. Opin. Struct. Biol.*, **13**, 40–48.
9. Remenyi, A., Scholer, H. R., and Wilmanns, M. 2004, Combinatorial control of gene expression, *Nat. Struct. Mol. Biol.*, **11**, 812–815.
10. Aerts, S., Van Loo, P., Thijs, G., Moreau, Y., and De Moor, B. 2003, Computational detection of cis-regulatory modules, *Bioinformatics*, **19** (Suppl 2), II5–II14.
11. Bejerano, G., Siepel, A. C., Kent, W. J., and Haussler, D. 2005, Computational screening of conserved genomic DNA in search of functional noncoding elements, *Nat. Methods*, **2**, 535–545.
12. Rajewsky, N., Vergassola, M., Gaul, U., and Siggia, E. D. 2002, Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo, *BMC Bioinformatics*, **3**, 30.
13. Matys, V., Fricke, E., Geffers, R., et al. 2003, TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
14. Sandelin, A. and Wasserman, W. W. 2004, Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics, *J. Mol. Biol.*, **338**, 207–215.
15. Xie, X., Lu, J., and Kulbokas, E. J. 2005, Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals, *Nature*, **434**, 338–345.
16. Schneider, T. D. and Stephens, R. M. 1990, Sequence logos: a new way to display consensus sequences, *Nucleic Acids Res.*, **18**, 6097–6100.
17. Antequera, F. 2003, Structure, function and evolution of CpG island promoters, *Cell Mol. Life. Sci.*, **60**, 1647–1658.
18. Waterston, R. H., Lindblad-Toh, K., Birney, E., et al. 2002, Initial sequencing and comparative analysis of the mouse genome, *Nature*, **420**, 520–562.
19. Suzuki, Y., Yamashita, R., Sugano, S., and Nakai, K. 2004, DBTSS, DataBase of Transcriptional Start Sites: progress report 2004, *Nucleic Acids Res.*, **32**, D78–D81.
20. 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.
21. Sharov, A. A., Piao, Y., Matoba, R., et al. 2003, Transcriptome analysis of mouse stem cells and early embryos, *PLoS Biol.*, **1**, E74.
22. Sharov, A. A., Dudekula, D. B., and Ko, M. S. 2005, Genome-wide assembly and analysis of alternative transcripts in mouse, *Genome. Res.*, **15**, 748–754.
23. Carninci, P., Kasukawa, T., Katayama, S., et al. 2005, The transcriptional landscape of the mammalian genome, *Science*, **309**, 1559–1563.
24. Gerhard, D. S., Wagner, L., Feingold, E. A., et al. 2004, The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC), *Genome Res.*, **14**, 2121–2127.
25. Okazaki, N., Kikuno, R., and Ohara, R., 2002, Prediction of the coding sequences of mouse homologues of KIAA gene: I. The complete nucleotide sequences of 100 mouse KIAA-homologous cDNAs identified by screening of terminal sequences of cDNA clones randomly sampled from size-fractionated libraries, *DNA Res.*, **9**, 179–188.
26. Pruitt, K. D., Tatusova, T., and Maglott, D. R. 2005, NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, *Nucleic Acids Res.*, **33**, D501–D504.
27. Birney, E., Andrews, D., Caccamo, M., et al. 2006, Ensembl 2006, *Nucleic Acids Res.*, **34**, D556–D561.
28. Marra, M., Hillier, L., Kucaba, T., et al. 1999, An encyclopedia of mouse genes, *Nat. Genet.*, **21**, 191–194.
29. Carter, M. G., Piao, Y., Dudekula, D. B., et al. 2003. The NIA cDNA project in mouse stem cells and early embryos, *C. R. Biol.*, **326**, 931–940.
30. Davuluri, R. V., Grosse, I., and Zhang, M. Q. 2001, Computational identification of promoters and first exons in the human genome, *Nat. Genet.*, **29**, 412–417.
31. Kent, W. J. 2002, BLAT–the BLAST-like alignment tool, *Genome Res.*, **12**, 656–664.
32. Benjamini, Y. and Hochberg, Y. 1995, Controlling the false discovery rate – a practical and powerful approach to multiple testing, *J. R Stat. Soc. B*, **57**, 289–300.
33. Kimura, K., Wakamatsu, A., Suzuki, Y., et al. 2006. Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes, *Genome Res.*, **16**, 55–65.
34. Carninci, P., Sandelin, A., Lenhard, B., et al. 2006. Genome-wide analysis of mammalian promoter architecture and evolution, *Nat. Genet.*, **38**, 626–635.
35. Suzuki, Y., Taira, H., Tsunoda, T., et al. 2001. Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites, *EMBO Rep.*, **2**, 388–393.
36. Zhang, Z. and Gerstein, M. 2003, Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements, *J. Biol.*, **2**, 11.
37. Kolbe, D., Taylor, J., Elnitski, L., et al. 2004, Regulatory potential scores from genome-wide three-way alignments of human, mouse, and rat, *Genome Res.*, **14**, 700–707.
38. Siepel, A., Bejerano, G., and Pedersen, J. S., 2005, Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes, *Genome Res.*, **15**, 1034–1050.
39. Blanchette, M., Bataille, A. R., Chen, X., et al. 2006, Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression, *Genome Res.*, **16**, 656–668.
40. Elnitski, L., Hardison, R. C., Li, J., et al. 2003, Distinguishing regulatory DNA from neutral sites, *Genome Res.*, **13**, 64–72.
41. Mulder, N. J., Apweiler, R., and Attwood, T. K. 2003, The InterPro Database, 2003 brings increased coverage and new features, *Nucleic Acids Res.*, **31**, 315–318.
42. Ashburner, M., Ball, C. A., Blake, J. A., et al. 2000, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nat. Genet.*, **25**, 25–29.
43. Yafe, A., Etzioni, S., Weisman-Shomer, P., and Fry, M. 2005, Formation and properties of hairpin and tetraplex structures of guanine-rich regulatory sequences of muscle-specific genes, *Nucleic Acids Res.*, **33**, 2887–2900.
44. Yamashita, R., Suzuki, Y., Sugano, S., and Nakai, K. 2005, Genome-wide analysis reveals strong correlation between CpG islands with nearby transcription start sites of genes and their tissue specificity, *Gene*, **350**, 129–136.

45. Saxonov, S., Berg, P., and Brutlag, D. L. 2006, A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters, *Proc. Natl Acad. Sci. USA*, **103**, 1412–1417.
46. Schug, J., Schuller, W. P., Kappen, C., Salbaum, J. M., Bucan, M., and Stoeckert, C. J.Jr 2005, Promoter features related to tissue specificity as measured by Shannon entropy, *Genome Biol.*, **6**, R33.
47. 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.
48. Vandepoele, K., Vlieghe, K., and Florquin, K. 2005, Genome-wide identification of potential plant E2F target genes, *Plant Physiol.*, **139**, 316–328.
49. Kreiman, G. 2004, Identification of sparsely distributed clusters of cis-regulatory elements in sets of co-expressed genes, *Nucleic Acids Res.*, **32**, 2889–2900.
50. Wasserman, W. W. and Fickett, J. W. 1998, Identification of regulatory regions which confer muscle-specific gene expression, *J. Mol. Biol.*, **278**, 167–181.
51. Mitsui, K., Tokuzawa, Y., Itoh, H., et al. 2003, The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells, *Cell*, **113**, 631–642.
52. Moon, H., Filippova, G., and Loukinov, D. 2005, CTCF is conserved from Drosophila to humans and confers enhancer blocking of the Fab-8 insulator, *EMBO Rep.*, **6**, 165–170.
53. Gu, P., Goodwin, B., Chung, A. C., et al. 2005, Orphan nuclear receptor LRH-1 is required to maintain Oct4 expression at the epiblast stage of embryonic development, *Mol. Cell. Biol.*, **25**, 3492–3505.
54. Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W. W., and Lenhard, B. 2004, JASPAR: an open-access database for eukaryotic transcription factor binding profiles, *Nucleic Acids Res.*, **32**, D91–D94.
55. Pritsker, M., Liu, Y. C., Beer, M. A., and Tavazoie, S. 2004, Whole-genome discovery of transcription factor binding sites by network-level conservation, *Genome Res.*, **14**, 99–108.
56. Hallikas, O., Palin, K., Sinjushina, N., et al. 2006, Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity, *Cell*, **124**, 47–59.
57. Down, T. A. and Hubbard, T. J. 2002, Computational detection and location of transcription start sites in mammalian genomic DNA, *Genome Res.*, **12**, 458–461.
58. Cartharius, K., Frech, K., Grote, K., et al. 2005, MatInspector and beyond: promoter analysis based on transcription factor binding sites, *Bioinformatics*, **21**, 2933–2942.
59. Long, L., Davidson, J. N., and Spear, B. T. 2004, Striking differences between the mouse and the human alpha-fetoprotein enhancers, *Genomics*, **83**, 694–705.
60. Vilaboa, N., Bermejo, R., Martinez, P., Bornstein, R., and Cales, C. 2004, A novel E2 box-GATA element modulates Cdc6 transcription during human cells polyploidization, *Nucleic Acids Res.*, **32**, 6454–6467.
61. Ohuchi, H., Yasue, A., Ono, K., et al. 2005, Identification of cis-element regulating expression of the mouse Fgf10 gene during inner ear development, *Dev. Dyn.*, **233**, 177–187.
62. Sasaki, H. and Hogan, B. L. 1996, Enhancer analysis of the mouse HNF-3 beta gene: regulatory elements for node/notochord and floor plate are independent and consist of multiple sub-elements, *Genes Cells*, **1**, 59–72.
63. Rojas, A., De Val, S., Heidt, A. B., Xu, S. M., Bristow, J., and Black, B. L. 2005, Gata4 expression in lateral mesoderm is downstream of BMP4 and is activated directly by Forkhead and GATA transcription factors through a distal enhancer element, *Development*, **132**, 3405–3417.
64. Ferretti, E., Marshall, H., Popperl, H., Maconochie, M., Krumlauf, R., and Blasi, F. 2000, Segmental expression of Hoxb2 in r4 requires two separate sites that integrate cooperative interactions between Prep1, Pbx and Hox proteins, *Development*, **127**, 155–166.
65. Wang, W. C., Anand, S., Powell, D. R., Pawashe, A. B., Amemiya, C. T., and Shashikant, C. S. 2004, Comparative cis-regulatory analyses identify new elements of the mouse Hoxc8 early enhancer, *J. Exp. Zool. B Mol. Dev. Evol.*, **302**, 436–445.
66. Niimi, T., Hayashi, Y., and Sekiguchi, K. 2003, Identification of an upstream enhancer in the mouse laminin alpha 1 gene defining its high level of expression in parietal endoderm cells, *J. Biol. Chem.*, **278**, 9332–9338.
67. Okumura-Nakanishi, S., Saito, M., Niwa, H., and Ishikawa, F. 2005, Oct-3/4 and Sox2 regulate Oct-3/4 gene in embryonic stem cells, *J. Biol. Chem.*, **280**, 5307–5317.
68. Petrovic, N., Black, T. A., Fabian, J. R., et al. 1996, Role of proximal promoter elements in regulation of renin gene transcription, *J. Biol. Chem.*, **271**, 22499–22505.
69. Okuno, Y., Huang, G., Rosenbauer, F., et al. 2005, Potential autoregulation of transcription factor PU.1 by an upstream regulatory element, *Mol. Cell. Biol.*, **25**, 2832–2845.
70. Perez-Mutul, J., Macchi, M., and Wasylyk, B. 1988, Mutational analysis of the contribution of sequence motifs within the IgH enhancer to tissue specific transcriptional activation, *Nucleic Acids Res.*, **16**, 6085–6096.
71. Saitsu, H., Komada, M., Suzuki, M., et al. 2005, Expression of the mouse Fgf15 gene is directly initiated by Sonic hedgehog signaling in the diencephalon and midbrain, *Dev. Dyn.*, **232**, 282–292.
72. Fomin, M., Nomokonova, N., and Arnold, H. H. 2004, Identification of a critical control element directing expression of the muscle-specific transcription factor MRF4 in the mouse embryo, *Dev. Biol.*, **272**, 498–509.
73. Catena, R., Tiveron, C., Ronchi, A., et al. 2004, Conserved POU binding DNA sites in the Sox2 upstream enhancer regulate gene expression in embryonic and neural stem cells, *J. Biol. Chem.*, **279**, 41846–41857.
74. Lele, K. M. and Wolgemuth, D. J. 2004, Distinct regions of the mouse cyclin A1 gene, Ccna1, confer male germ-cell specific expression and enhancer function, *Biol. Reprod.*, **71**, 1340–1347.
75. Shimizu, M., Furuya, S., Shinoda, Y., et al. 2004, Functional analysis of mouse 3-phosphoglycerate dehydrogenase (Phgdh) gene promoter in developing brain, *J. Neurosci. Res.*, **76**, 623–632.
76. Ijichi, N., Tsujimoto, N., Iwaki, T., Fukumaki, Y., and Iwaki, A. 2004, Distal Sox binding elements of the alphaB-crystallin gene show lens enhancer activity in transgenic mouse embryos, *J. Biochem. (Tokyo)*, **135**, 413–420.
77. Li, T., Lu, Z., and Lu, L. 2004, Regulation of eye development by transcription control of CCCTC binding factor (CTCF), *J. Biol. Chem.*, **279**, 27575–27583.

78. Ponticos, M., Abraham, D., Alexakis, C., et al. 2004, Col1a2 enhancer regulates collagen activity during development and in adult tissue repair, *Matrix Biol.*, **22**, 619–628.
79. Brown, C. O., 3rd, Chi, X., Garcia-Gras, E., Shirai, M., Feng, X. H., and Schwartz, R. J. 2004, The cardiac determination factor, Nkx2–5, is activated by mutual cofactors GATA-4 and Smad1/4 via a novel upstream enhancer, *J. Biol. Chem.*, **279**, 10659–10669.
80. Mortlock, D. P., Guenther, C., and Kingsley, D. M. 2003, A general approach for identifying distant regulatory elements applied to the Gdf6 gene, *Genome Res.*, **13**, 2069–2081.
81. Ishida, C., Ura, K., Hirao, A., et al. 2003, Genomic organization and promoter analysis of the Dnmt3b gene, *Gene*, **310**, 151–159.
82. Spitz, F., Gonzalez, F., and Duboule, D. 2003, A global control region defines a chromosomal regulatory landscape containing the HoxD cluster, *Cell*, **113**, 405–417.
83. Yokouchi, K., Ito, M., Nishino, K., et al. 2003, Stage-specific regulatory element of mouse Sry gene, *Mol. Reprod. Dev.*, **64**, 389–396.
84. Leighton, P. A., Saam, J. R., Ingram, R. S., Stewart, C. L., and Tilghman, S. M. 1995, An enhancer deletion affects both H19 and Igf2 expression, *Genes Dev.*, **9**, 2079–2089.
85. Szabo, P., Tang, S. H., Rentsendorj, A., Pfeifer, G. P., and Mann, J. R. 2000, Maternal-specific footprints at putative CTCF sites in the H19 imprinting control region give evidence for insulator function, *Curr. Biol.*, **10**, 607–610.
86. Bagheri-Fam, S., Barrionuevo, F., Dohrmann, U., et al. 2006, Long-range upstream and downstream enhancers control distinct subsets of the complex spatiotemporal Sox9 expression pattern, *Dev. Biol.*, **291**, 382–397.
87. Kurokawa, D., Kiyonari, H., Nakayama, R., Kimura-Yoshida, C., Matsuo, I., and Aizawa, S. 2004, Regulation of Otx2 expression and its functions in mouse forebrain and midbrain, *Development*, **131**, 3319–3331.
88. Hadchouel, J., Carvajal, J. J., Daubas, P., et al. 2003, Analysis of a key regulatory region upstream of the Myf5 gene reveals multiple phases of myogenesis, orchestrated at each site by a combination of elements dispersed throughout the locus, *Development*, **130**, 3415–3426.