High-resolution human core-promoter prediction with CoreBoost_HM

Xiaowo Wang,[1,2,3] Zhenyu Xuan,[2,3] Xiaoyue Zhao,[2] Yanda Li,[1] and Michael Q. Zhang[2,4]

[1]*MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST/Department of Automation, Tsinghua University, Beijing 100084, China;* [2]*Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA*

Correctly locating the gene transcription start site and the core-promoter is important for understanding transcriptional regulation mechanism. Here we have integrated specific genome-wide histone modification and DNA sequence features together to predict RNA polymerase II core-promoters in the human genome. Our new predictor CoreBoost_HM outperforms existing promoter prediction algorithms by providing significantly higher sensitivity and specificity at high resolution. We demonstrated that even though the histone modification data used in this study are from a specific cell type (CD4+ T-cell), our method can be used to identify both active and repressed promoters. We have applied it to search the upstream regions of microRNA genes, and show that CoreBoost_HM can accurately identify the known promoters of the intergenic microRNAs. We also identified a few intronic microRNAs that may have their own promoters. This result suggests that our new method can help to identify and characterize the core-promoters of both coding and noncoding genes.

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

Transcription initiation is a key step in the regulation of gene expression. During this process, transcription factors bind promoter region of a gene in a sequence-specific manner and recruit the RNA polymerase to form an active initiation complex around the transcription start site (TSS). The promoter region is typically divided into three parts: (1) a core-promoter that is ~100 bp long around the TSS containing binding sites for RNA polymerase II (Pol II) and general transcription factors; (2) a proximal promoter of several hundred base pairs long containing primary specific regulatory elements located immediately upstream of the core-promoter; and (3) a distal promoter up to thousands of base pairs long providing additional regulatory information. As it contains primary information to control gene transcription, it is a fundamental step to identify the core-promoter in study of gene expression patterns and constructing gene transcription networks.

It has been noticed that promoter regions contain characteristic features that can be used to distinguish them from other parts of the genome. These features may be grouped into two types: one is on small-scale, e.g., TATA-box, GC-box, CAAT-box, and Inr; the other is on larger-scale, such as CpG island, *k*mer frequency, density of transcription factor binding sites, nucleosome binding, and chromatin modifications (Bajic et al. 2004; Zhang 2007). Recently, the large-scale DNA structural features have successfully been used to improve the promoter predictions (Abeel et al. 2008). Accordingly, a two-step approach has been proposed for TSS identification (Zhang 1998): First, use the large-scale features to roughly locate a promoter in a 1- to 2-kb region (low resolution), then use the small-scale features to refine the prediction into a 100-bp core-promoter region (high resolution). In the past 10 yr, dozens of computational methods have been developed to predict core-promoters (for review, see Bajic et al. 2004, 2006; Zhang 2007). However, it is still a challenging problem to identify functional TSS accurately at high resolution, especially for the non-CpG-related promoters. In our previous study, we developed a program called CoreBoost (Zhao et al. 2007), which applied a boosting technique with stumps to predict core-promoters, using both the small-scale and large-scale DNA sequence features. This program shows significant advantage for high-resolution prediction than other popular promoter predictors (Zhao et al. 2007).

Up to now, almost all of the existing core-promoter prediction methods are based on DNA sequence features, few of them has incorporated the epigenetic information. Eukaryotic DNA is packaged into chromatin structure. The basic unit of chromatin is the nucleosome that is composed of an octamer of four core histones (H2A, H2B, H3, and H4) wrapped by 147-bp-long DNA. The modifications of the histones are found to be associated with transcription initiation and elongation (for review, see Kouzarides 2007). For example, nucleosomes with H3 lysine 4 tri-methylation (H3K4me3) are enriched in the promoter regions, and H3 lysine 36 tri-methylation (H3K36me3) occurs at nucleosomes covering primary transcripts of actively expressed genes (Mikkelsen et al. 2007). Recently, taking advantage of next-generation sequencing technology, researchers developed the chromatin immunoprecipitation followed by sequencing (ChIP-seq) technology and obtained high-throughput genome-wide profiles of histone modification patterns in mammalian cells (Barski et al. 2007; Mikkelsen et al. 2007). These high-resolution histone modification profiling data shed new light on studying regulation mechanism of gene transcription.

In this work, we systematically analyzed different chromatin features for promoter prediction and integrated the specific histone modification profiles and the DNA sequence features together to predict human Pol II core-promoters. Our new predictor CoreBoost_HM (CoreBoost with histone modification features) significantly outperforms the best existing methods by providing much higher sensitivity and specificity for high-resolution TSS predictions. Though the histone modification data used in this

[3]These authors contributed equally to this work.
[4]Corresponding author.
E-mail mzhang@cshl.org; fax (516) 367-8461.

Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.081638.108.

study are from a specific cell type (CD4+ T-cell), surprisingly CoreBoost_HM can nevertheless identify both active and repressed promoters accurately. As a novel application, we also used CoreBoost_HM to search the upstream regions of microRNA (miRNA) genes, a class of short endogenous noncoding RNAs that play important regulatory roles (Bartel 2004). CoreBoost_HM gives a good prediction on the known intergenic miRNA promoters. And interestingly, we also found that some of intronic miRNAs appear to have their own promoters that are different from the promoters of their host genes. This result suggests that CoreBoost_HM can help to identify and characterize the core-promoters of both coding and non-coding genes.

Results and Discussion

Histone modification signal around CpG- and non-CpG-related promoters

Previous studies suggest that CpG- and non-CpG-related promoters have distinct sequence features and should be treated separately (Davuluri et al. 2001; Bajic et al. 2002; Saxonov et al. 2006). Recent results in human T cells (Roh et al. 2006) and mouse ES cells (Mikkelsen et al. 2007) also suggest that the histone modification profiles around these two types of promoters are different. As shown in Figure 1A, for the CpG-related promoters, H3K4me3 signal has two major peaks located upstream and downstream of the TSS (where the trough presumably reflects the nucleosome depletion at the core-promoter) (Schones et al. 2008), respectively. But for the non-CpG-related promoters, not only is the upstream peak of H3K4me3 significantly lower but also the overall modification signal intensity is much weaker than that of CpG-related promoters. Because the CpG-related promoters compose the majority of the known promoters (~72%) (Saxonov et al. 2006) and have stronger signal intensity, if these two types of promoters are mixed together, the overall patterns will be dominated by the CpG-related promoters (Barski et al. 2007). The differences between CpG- and non-CpG-related promoters were also observed for some other histone markers (Fig. 1B; Supplemental Fig. 1). Thus, similarly as with the strategy of CoreBoost, we analyzed CpG and non-CpG promoters separately.

Use histone markers for promoter prediction

Though it is widely believed that epigenomic features (like histone modifications) can provide an extra layer of information beyond DNA sequence features for correctly locating gene promoters, up to now, few computational methods have been developed to extract such information systematically. Recently, Heintzman et al. (2007) analyzed the ChIP with microarray hybridization (ChIP-chip) data of several histone acetylation and methylation markers within the ENCODE region and found that H3K4me3 profile can be used to identify active promoters. However, they only tested five histone markers and just used two of them for low-resolution predictions in the ENCODE region. In this study, we explored the contribution of different histone modification features for promoter prediction. The whole-genome histone modification profiles of 20 histone methylations, 18 acetylations, and one histone variant H2A.Z in human CD4+ T-cells were extracted from recent ChIP-seq experiment data sets (Barski et al. 2007; Wang et al. 2008b). A training set of 4263 CpG-related and 1683 non-CpG-related human Pol II core-promoters are used in this analysis. This set is constructed based on the known TSS in the Eukaryotic Promoter Database (EPD) (Cavin Perier et al. 1998) and DBTSS (Wakaguri et al. 2008) databases, with expression profiles of related genes in human CD4+ T-cell from the GNF gene expression atlas (Su et al. 2004). Our preliminary analysis suggested that for each histone modification signal, both their shape and absolute intensity around the TSS may contribute positively to the prediction. So we encoded each type of histone modification signal with two features: One is the Pearson correlation coefficient between the vector of local modification profile and the vector of empirical average pattern of this modification around all promoters in the training set, which is used to capture the similarity of the "shape" signal; the other is the dot product of these two vectors, which reflects the weighted intensity of this modification around a TSS. Finally, LogitBoost with stumps (Friedman et al. 2000) was used for the classification, which can select informative features and use an assembly of weak classifiers to build a strong classifier. To understand the contribution of each type of modifications, we conducted separate computational experiments using the information from (1) the 20 histone methylations, (2) the 18 histone acetylations, and (3) all the 39 histone markers, respectively. A 10-fold cross-validation was performed, and the performance was evaluated by searching the 2.4-kb region centered at the TSS of each promoter in the training set.

As shown in Figure 2, A and C, acetylation and methylation markers do contribute to promoter prediction, especially for the CpG-related promoters. Table 1 lists the top histone markers contribute to boosting classifiers. Most of these top features are consistently selected during the cross-validation (Supplemental Table S1). As expected, H3K4me3 appears to be the most important histone feature for promoter prediction. But somewhat surprisingly, acetylations only provide subtle additional contribution when we put all the available histone markers together. Besides, we also noticed that though the well-known promoter associated markers like H3K4me3 and H2A.Z contribute

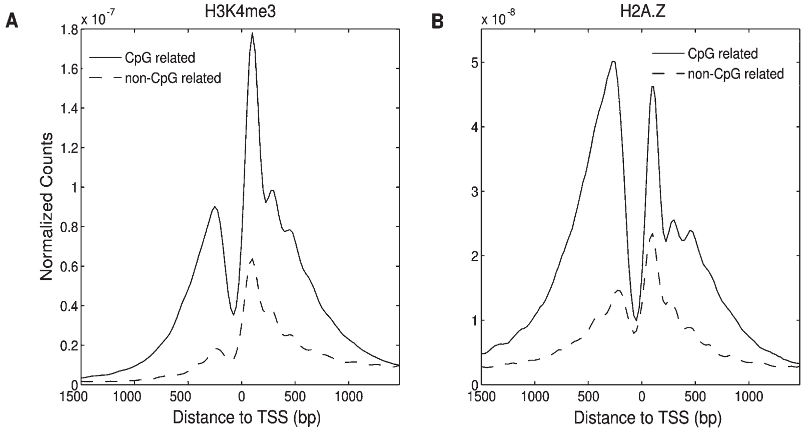

Figure 1. Histone modification profiles of H3K4me3 and H2A.Z around gene transcription start sites H3K4me3 (*A*) and H2A.Z (*B*) are two of the top histone markers that contribute to core-promoter prediction. CpG-related and non-CpG-related promoters have significantly different histone modification profiles around the TSS. Normalized counts are reads counts normalized by the total number of reads in the data set.

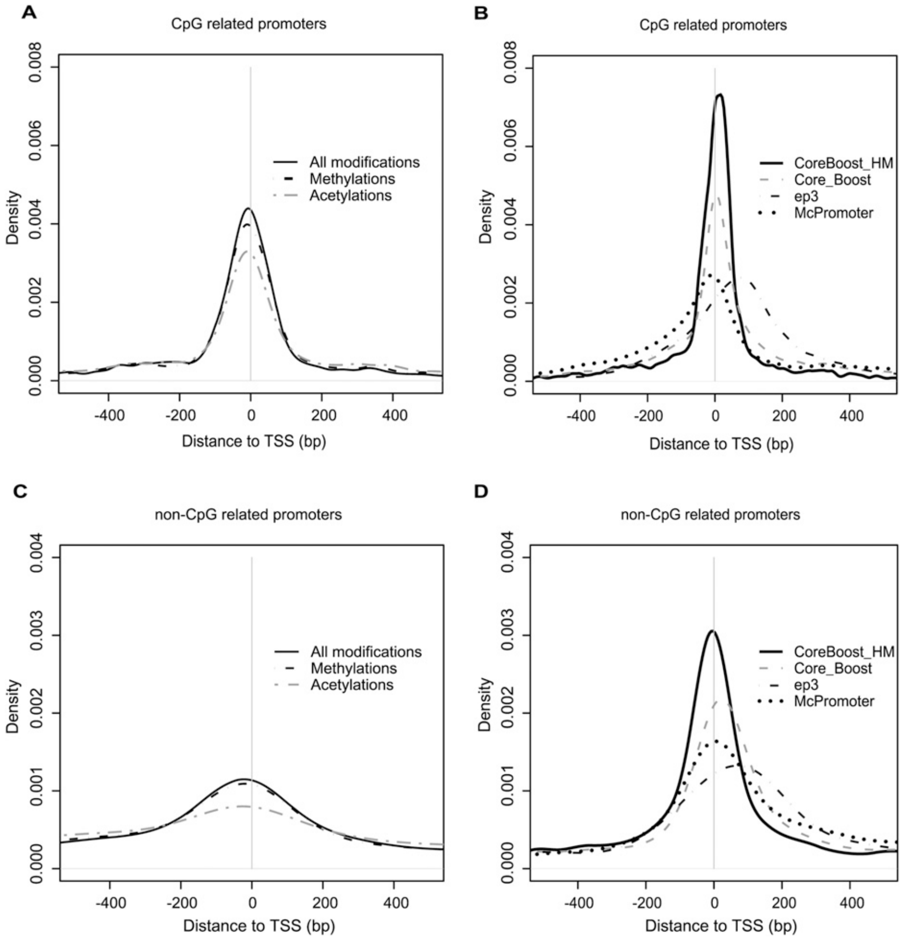

Figure 2. Density plot of the relative distance from the positions with maximal prediction scores to the annotated TSS. Chromatin features contribute to the prediction of both CpG-related (*A*) and non-CpG-related (*C*) promoters. (*B*,*D*) Further combining chromatin signatures with sequence features can greatly improve the prediction accuracy.

greatly to both the CpG-related and non-CpG-related promoter predictors, these two predictors may be combined with different histone features separately to achieve optimal performance. For example, the H3K4me2 is an important complement to H3K4me3 for the prediction of CpG-related promoters, but it seems to be redundant for the prediction of non-CpG-related promoters. This observation supports the notion that the CpG-related and non-CpG-related promoters may be subject to distinct mode of regulation (Mikkelsen et al. 2007). To further study the contribution from the top features, we also investigated the prediction power when only the top one, three, five, and 10 histone markers were used in each group. We found that using the features from a top few chromatin markers are sufficient for lower-resolution (i.e., large-scale, ~500-bp) prediction (Supplemental Tables S2, S3), but adding additional features can further improve the accuracy at high resolution (i.e., fine scale, ~50 bp), especially for CpG-related promoters. However, not all histone markers are informative for the classification, using about the top 20 markers is sufficient for the prediction, and adding further features do not show significant improvement.

Many histone markers associates with gene expression, and their profiling signals are very different between active and silent promoters (Barski et al. 2007). In order to study the correlation between gene expression and promoter prediction power of chromatin signatures, we further split the CpG-related and non-CpG-related promoters in the training set into two equal size groups, respectively, according to the gene expression level—(1) top one half of genes with the highest expression level (active genes) and (2) bottom one half of genes with the lowest expression level (low express/silenced genes)—and used these subsets for the boosting analysis. Surprisingly, we noticed that although the intensities of many histone markers are very different between active and silenced promoters, the top features selected by boosting classifier are quite similar (Supplemental Table S4), which suggests that, in general, the difference of histone modification patterns

Table 1. Top histone markers contribute to promoter prediction

Methylations		Acetylations		All histone markers	
CpG[a]	Non-CpG[a]	CpG[a]	Non-CpG[a]	CpG[a]	Non-CpG[a]
H3K4me3	H3K4me3	H3K18ac	H3K27ac	H3K4me3	H3K4me3
H3K4me1	H3K79me3	H2AK5ac	H4K91ac	H3K4me1	H2A.Z
H3K4me2	H3K79me1	H4K91ac	H3K18ac	H3K4me2	H3K79me3
H3K79me3	H4R3me2	H3K23ac	H4K16ac	H3K79me3	H3K27ac
H4K20me1	H4K20me3	H4K5ac	H3K9ac	H4K20me1	H3K79me1
H3K27me1	H3K27me2	H3K14ac	H2BK5ac	H2AK5ac	H4R3me2
H3K9me2	H3K36me3	H3K9ac	H4K8ac	H2A.Z	H3K27me3
H4K20me3	H3K27me3	H2BK12ac	H2BK20ac	H4K91ac	H3K18ac
H3K36me1	H3K4me1	H2AK9ac	H2BK12ac	H3K23ac	H4K91ac
H3K9me1	H3K9me2	H3K36ac	H3K36ac	H3K27me1	H3K27me2

These markers were sorted according to the order they were selected by the boosting classifier.
[a]For CpG-related and non-CpG-related promoters, respectively.

between active and repressed promoter is smaller than that between different types of promoters (CpG and non-CpG). What's more, prediction results suggest that although these histone markers contribute more for the prediction of active promoters, they can also be used to predict inactive ones with remarkable accuracy (Supplemental Table S5). This result is supported by the recent notion that some of the histone markers like H3K4me3 and H2A.Z associate with both active and silenced promoters (for review, see Schones and Zhao 2008), and it also indicates that these histone features can be used to design general promoter predictors rather than specific ones (restrict only to active promoters).

CoreBoost with histone modification features

Next, we integrated chromatin signatures with DNA sequence features together to design a high-resolution human Pol II core-promoter prediction algorithm. We followed the framework of CoreBoost, and the inputs to our new predictor CoreBoost_HM include both genomic sequences and the histone modification profiles. The sequence features were calculated according to our previous work (Zhao et al. 2007). These features were mainly extracted from a 300-bp region (−250 to 50) around the TSS, including the core-promoter elements score, transcription factor binding site (TFBS) density, Markovian log-likelihood ratio scores, and *k*mer frequencies. In addition, we also considered some long range DNA energy and flexibility properties (Packer et al. 2000) in a larger region expanding from upstream −1300 to downstream +1300 bp relative to the TSS. These long-range features (calculated based on tetra nucleotide properties) were found to be very useful for the prediction of non-CpG-related promoters (Zhao et al. 2007). Most of the top features used in CoreBoost_HM (Supplemental Table S6) were also selected by the predictors trained by sequence or histone modification features alone, which partially indicates that histone modification features can provide additional information beyond currently used sequence features. The top contributors for CpG-related predictor are log-likelihood ratios from third-order Markov chain, Profile correlation score of H3K4me3 and H3K4me2, GC box score, etc.; and the top contributors for the non-CpG model are the correlation score of DNA energy profile, the profile correlation score of H3K4me3, log-likelihood ratios from third-order Markov chain, TATA score, and Inr score, etc. In total, 18 and 17 histone markers are used in the CpG-related and non-CpG-related classifiers, respectively, but some of them only contribute marginally to the classification (Supplemental Table S6). The detailed description of these features can be found in the Methods section on feature extraction and also in Zhao et al. (2007).

Comparison with other methods

We compared the performance of CoreBoost_HM with the Boosting predictor using only the histone modification features (BHMF), original CoreBoost, and other two free and publicly available high-performance promoter predictors, McPromoter (Ohler et al. 2001) and EP3 (Abeel et al. 2008). McPromoter is a probabilistic promoter predictor that uses a neural network to combine the sequence features and structural profiles in prediction. EP3 is one of the latest core-promoter prediction algorithms based on large-scale DNA structural features. Abeel et al. (2008) compared EP3 with a number of promoter prediction programs and demonstrated that it outperforms most of the other state-of-the-art algorithms.

We first evaluated the performance of each predictor by searching the 2.4-kb region centered at the TSS of each promoter in the training set, and the prediction power of CoreBoost_HM and BHMF was estimated by a 10-fold cross-validation. Figure 2 shows the density plots of the relative distance from the positions with the maximal prediction score to the true TSS. About 58.6% and 35.9% CoreBoost_HM maximal-score predictions locate within 50 bp of the annotated TSS for CpG-related and non-CpG-related promoters, respectively, which are significantly higher than that only using sequence (original CoreBoost, 38.9% and 32.0%) or histone features (BHMF, 47.8% and 20.0%) and much better than that of McPromoter (27.5% and 23.8%) and EP3 (22.9% and 14.5%). Figure 3 shows the plots of positive prediction values (PPVs) versus sensitivity for these prediction programs at the precision request of 50 bp, which further confirm the power of CoreBoost_HM for high-resolution predictions. Supplemental Table S7 shows the performance evaluated at different resolution on the training set.

To further evaluate the performance on larger genomic regions, we tested these programs on an independent test set, which is composed of the upstream 5-kb to downstream 5-kb region of 1642 nonoverlap gene promoters, each containing one or multiple TSS. We adopted the *F*-score (Abeel et al. 2008) to measure the performance at precision requisition of 50, 200, and 500 bp. We varied different thresholds of prediction score for each program and chose the threshold with maximum *F*-score. As shown in Table 2, CoreBoost_HM significantly outperforms BHMF and other three sequence based programs at high resolution. When allowing maximum distance from true TSS as 50 bp, CoreBoost_HM achieved an *F*-score of 0.32, which is about or more than 1.5 times larger than that of BHMF (0.23), CoreBoost (0.20), McPromoter (0.18), and EP3 (0.11). For low-resolution prediction (>500 bp), as CoreBoost_HM (and its ancestor CoreBoost) is specially designed and optimized for high-resolution predictions, its

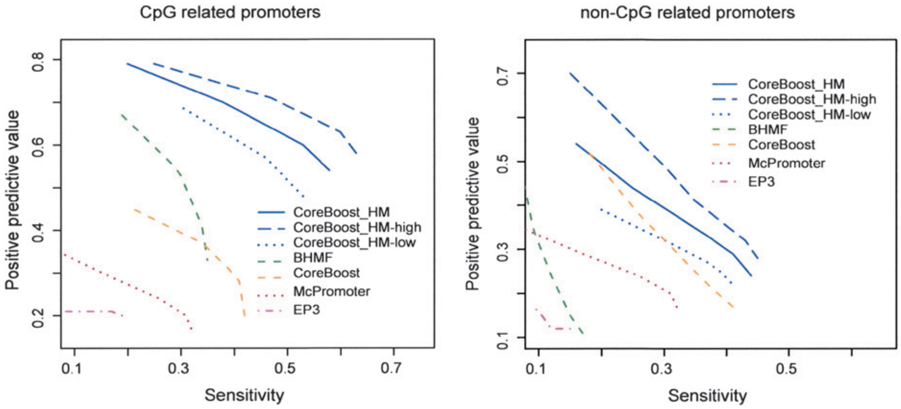

Figure 3. Positive predictive values (PPV) versus sensitivity for CpG-related and non-CpG-related promoters. The prediction results within 500 bp are clustered for each program. Only a prediction locating within 50 bp of the annotated TSS is regarded as a true-positive hit. The blue dashed and blue dotted lines indicate CoreBoost_HM performance on the promoters of highly expressed genes and low express genes, respectively.

Table 2. Performance evaluation on test set

	Maximum allowed distance from true TSS								
	50 bp			200 bp			500 bp		
Program	Sensitivity	PPV	F	Sensitivity	PPV	F	Sensitivity	PPV	F
CoreBoost_HM	0.26	0.43	0.32	0.45	0.56	0.50	0.57	0.72	0.64
CoreBoost_HM-high[a]	0.34	0.37	0.36	0.50	0.54	0.52	0.66	0.65	0.66
CoreBoost_HM-low[b]	0.26	0.29	0.27	0.42	0.48	0.45	0.59	0.62	0.60
BHMF	0.20	0.28	0.23	0.38	0.53	0.44	0.52	0.72	0.60
CoreBoost	0.19	0.20	0.20	0.38	0.39	0.38	0.55	0.56	0.55
McPromoter	0.17	0.19	0.18	0.43	0.32	0.37	0.67	0.54	0.60
EP3	0.12	0.11	0.11	0.40	0.37	0.38	0.62	0.61	0.62

[a]CoreBoost_HM performance on the promoters of highly expressed genes.
[b]CoreBoost_HM performance on the promoters of low expressed genes.
PPV, positive predictive value; F, harmonic average of sensitivity and PPV.

advantage becomes less pronounced, although it still slightly outperforms the state-of-the-art predictors such as EP3. We noticed that about one-quarter of CoreBoost_HM predictions on the test set are more than 500 bp away from the DBTSS or EPD annotations. Our further analysis revealed that ~22% and 44% of these "false-positive" predictions according to DBTSS or EPD annotation are supported by CAGE (Carninci et al. 2006) tag clusters at the 50-bp and 500-bp resolution, respectively. This result suggests that a considerable portion of CoreBoost_HM false-positive predictions are probably true alternative TSS. Performance evaluation using the combined annotation of DBTSS, EPD, and CAGE tag clusters is shown in Supplemental Table S8. In addition, Supplemental Table S9 shows the false-positive rate of CoreBoost_HM at different parameter settings estimated by searching randomly selected intronic regions that do not overlap with any DBTSS annotation, CAGE tag clusters, or Pol II peaks.

Gene expression level versus prediction accuracy

As the histone modification data in this study are from one cell type (CD4+ T-cell), it is essential to know whether our new predictor can be used for general promoter prediction or only the active promoters in this specific cell type. We further analyzed the CoreBoost_HM prediction results on both training set (based on cross validation) and test set regarding to downstream gene expression level. The TSS in each set were split into two equally sized subsets according to downstream gene expression level in CD4+ T-cell, with (1) top one half of genes with the highest expression level (active genes) and (2) bottom one half of genes with the lowest expression level (low express/silenced genes). As shown in both the training (Fig. 3; Supplemental Table S5) and the test set (Table 2), gene expression level does influence CoreBoost_HM prediction accuracy. However, comparing with the original CoreBoost, we found that the histone modification signals could still help to improve the performance for the prediction of less-active or repressed promoters, especially for the CpG-related ones.

Identification of miRNA gene core-promoters

miRNAs are short endogenous small RNA molecules that play essential regulatory roles in diverse organisms (for review, see Bartel 2004). Similar as pre-mRNAs, primary transcripts of intergenic miRNAs (called pri-miRNAs) also have 5′ cap structure and 3′ polyA tail (Cai et al. 2004; Lee et al. 2004). Though some of the pri-miRNAs are reported as Pol III–mediated transcription products (Borchert et al. 2006), the majority of them are believed to be transcribed by Pol II (Cai et al. 2004; Lee et al. 2004; Zhou et al. 2007). Up to now, only a few miRNA promoters have been experimentally characterized. In this study, we analyzed the upstream region of miRNAs using CoreBoost_HM. Figure 4, A through F, shows the predictions on the six intergenic miRNA promoters that are the only known ones in human. mir-23a cluster is the first intergenic pri-miRNA transcript experimentally characterized in human cells. Lee et al. (2004) demonstrated that these miRNAs share a common promoter within ~600 bp upstream of the cluster. CoreBoost_HM correctly predicts such a TSS at ~400 bp upstream of this miRNA cluster, which is also supported by an independent Pol II signal according to Pol II ChIP-seq data (Fig. 4A). The experimentally located promoter of mir-21 in HeLa cells is located at ~2.45 kb upstream of the pre-miRNA (Cai et al. 2004). CoreBoost_HM predicts a promoter at the upstream ~2.94-kb position, which is again supported by the Pol II signal (Fig. 4B). Additionally we also noticed that there are two weak predictions located at the upstream ~2-kb and ~3.4-kb position, respectively, both of them supported by either CAGE (Carninci et al. 2006) tags or the Pol II signal. We suspected that there may exist multiple alternative TSS for this miRNA gene. mir-17 cluster and mir-34a have been intensively studied for their potential important role in tumorigenesis, and both their core-promoters have been determined by RACE experiments recently (Chang et al. 2007; Woods et al. 2007). As shown in Figure 4C, CoreBoost_HM gives a strong CpG-related prediction at the upstream ~2.95-kb position relative to mir-17 precursor, less than 50 bp apart from the experimentally identified TSS. In the case of mir-34a, though the promoter is located at more than 30 kb away from the pre-miRNA, CoreBoost_HM precisely identify this CpG-related promoter. In addition, it predicted a putative alternative TSS which is also supported by CAGE tag clusters at the 600 bp further upstream of the known one (Fig. 4D). As another example, mir-223 has been characterized to have a core-promoter highly conserved between human and mouse (Fukao et al. 2007). CoreBoost_HM correctly gives a prediction located at the upstream ~3.4-kb position, which is only a few bases apart from the true TSS (Fig. 4E). The mir-371 cluster, which is specially expressed in embryonic stem cells and evolves fast, has been annotated to have a promoter located at the upstream ~700-bp position by a comparative genomic analysis (Houbaviy et al. 2005). CoreBoost_HM prediction score profile has two peaks: One is between mir-371 and mir-372, and the other is at the upstream ~2.1-kb

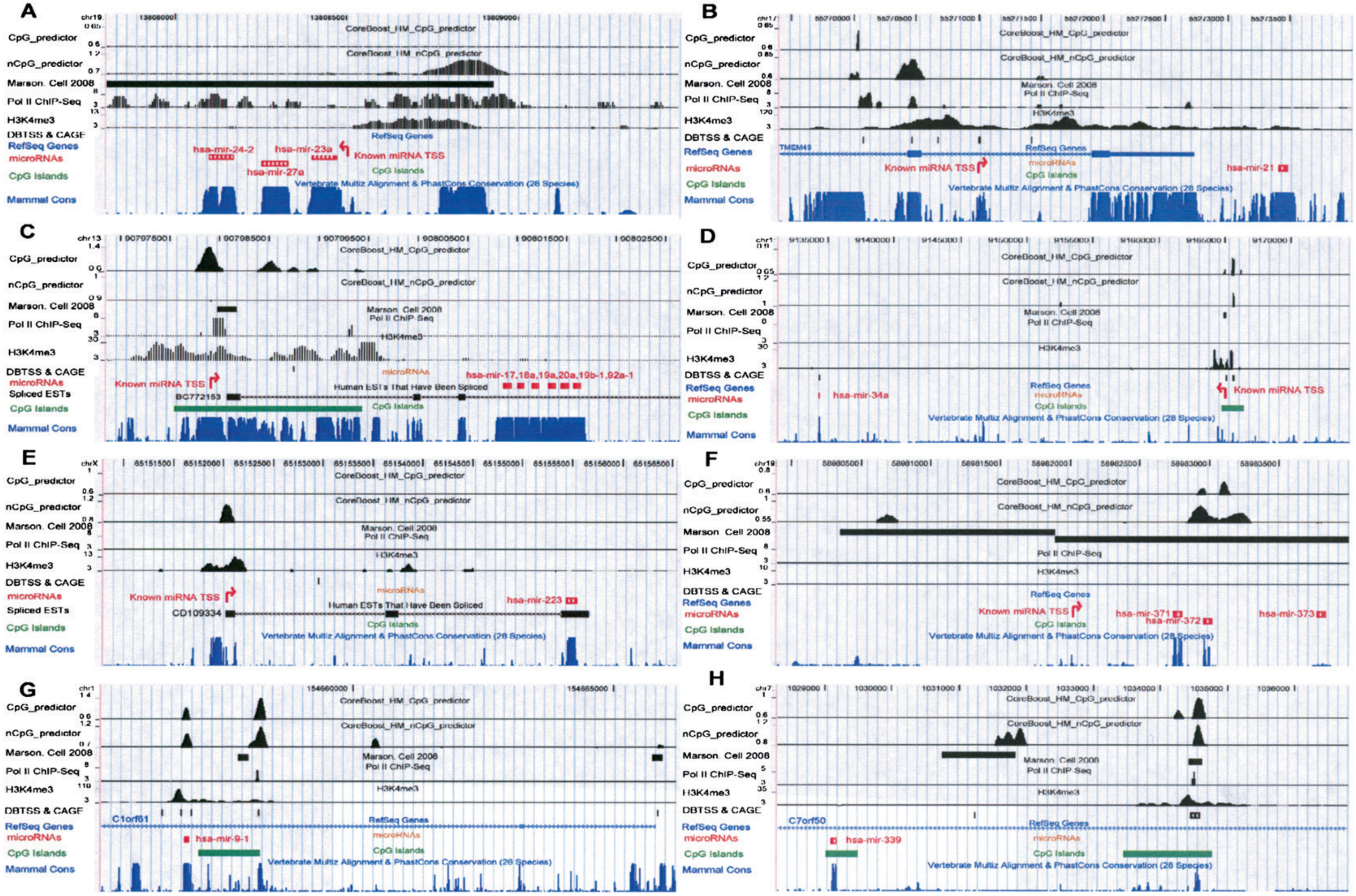

Figure 4. (*A–F*) CoreBoost_HM perform well on the known intergenic miRNA core-promoters. (*G*,*H*) CoreBoost_HM search suggests that some of the intronic miRNA genes may have their own promoters that are independent from the annotated promoters of their host genes. The third (*top*) track in each subplot shows the predictions given by Marson et al. (2008).

position (Fig. 4F). No other support (like CAGE or Pol II signal) is available for either the CoreBoost_HM predictions or the TSS inferred from its mouse orthologs. In total, CoreBoost_HM identified three of the six known miRNA promoters within 50 bp and the other two within 500 bp. Hence, although the majority of these known miRNA promoters are non-CpG-related and are not highly active in T cells according to the miRNA expression atlas (Landgraf et al. 2007), CoreBoost_HM is still able to make a reasonable prediction. For comparison, EP3 and the recently published miRNA promoter predictor CoVote (Zhou et al. 2007) can identify four and two of these six known promoters at 500-bp resolution, respectively (Supplemental Table S10).

Encouraged by these results, we used CoreBoost_HM to analyze the upstream 10 kb of the other intergenic miRNAs. Using a stringent threshold of 0.7 and 1.1 for CpG and non-CpG predictors, respectively, CoreBoost_HM predicted 145 putative TSS for 82 intergenic miRNAs or miRNA clusters, covering 156 pre-miRNAs. This parameter setting corresponds to an estimated sensitivity of 28% and 57% at 50-bp and 500-bp resolution, respectively and a false-positive number of ~42 (Supplemental Table S12). About 45% of these predictions are supported by either Pol II signal or CAGE tag clusters within 500 bp (summarized in Supplemental Table S11), and many of them show significantly higher average PhastCons conservation scores (Siepel et al. 2005) than the randomly selected control regions (Supplemental Fig. S2). Supplemental Table S12 lists CoreBoost_HM predictions on intergenic miRNAs with varying thresholds.

Experimental studies on several intronic miRNAs have suggested that such miRNAs are typically transcribed together with their host genes (Rodriguez et al. 2004). However, CoreBoost_HM predicts that a few intronic miRNAs may have their own promoters (Supplemental Table S13). As shown in Figure 4G, mir-9-1 is located in the second intron of its host gene *C1orf61*. CoreBoost_HM predicts a CpG-related promoter at 1.4 kb upstream of this miRNA. This predicted promoter is supported by both Pol II ChIP-seq and spliced expressed sequence tag (EST) data. Figure 4H shows another example, the mir-339, which is located in a long intron of the host gene *C7orf50*. A non-CpG-related and a CpG-related promoter were predicted at upstream 2.9 kb and 5.4 kb of mir-339, respectively, with the latter one supported by CAGE tags, spliced ESTs, and Pol II ChIP-seq data. These results suggest that some of the intergenic miRNAs are likely to be transcribed independently from their host genes.

During the revision of this manuscript, Marson et al. (2008) published their predicted human and mouse miRNA promoters. They analyzed the upstream 250 kb of each miRNA by incorporating H3K4me3 signature, conservation between human and murine, CpG islands, and ESTs among other features and provided low-resolution (~1 kb) predictions for 80% of the human miRNAs. We compared their predictions located within upstream 10 kb of the intergenic miRNAs with CoreBoost_HM predictions. In total, they predicted 77 putative promoters in these regions, 46 (60%) of them fall within 500 bp of 63 (43%) CoreBoost_HM high-confident predictions. Focusing on the known examples, as shown in Figure 4, A through F, their predictions covered four known promoters but missed the promoters of mir-21 and mir-223, which were predicted correctly by CoreBoost_HM. Interestingly, we also noticed that CoreBoost_HM predictions on the two intronic miRNAs shown in Figure 4, G and H, are also supported by their predictions. Although results only on the few examples may not be generalized, both methods unambiguously demonstrate that incorporating chromatin signatures can greatly facilitate the identification of miRNA promoters. Further analysis combining the advantage of both methods can produce a more accurate TSS map for the annotation of human miRNA genes.

Conclusion

In this work, we have shown that incorporating the genome-wide histone modification profiling data with DNA sequence features can greatly improve the accuracy of the Pol II core-promoter prediction in the human genome. We have systematically analyzed the prediction power of different chromatin signatures and developed a state-of-the-art core-promoter predictor called CoreBoost_HM, which outperforms several existing algorithms. We demonstrated that although the histone modification features are more informative for the active promoters, they can also help to find the less-active or even silenced promoters. Taking advantage of next-generation sequencing technology, we expect more and more histone modification ChIP-seq data sets to be produced from different cell types in the near future. These data will help to further improve the prediction accuracy and annotate the tissue-specific promoters. Another interesting observation in this study is that although many histone markers show characteristic patterns in the promoter region, some of them are redundant for promoter prediction. This work can be used as a guide for the design of ChIP experiments aiming to identify gene promoters.

Similar to CoreBoost, CoreBoost_HM is designed for fine-scale core-promoter localization problem but is not intended for genome-wide searching. And it should be noticed that, the non-CpG predictor contribute very little to the overall performance measured by *F*-score when searching large genomic regions. Though this separate classifier does provide more accurate predictions for non-CpG-related promoters, it would introduce more false positives as non-CpG-related promoters typically have weaker signals and only represent a minority in the genome. In practice, we recommend using some prior information to first identify a reasonable search region of ~2–5 kb and then applying CoreBoost_HM, especially for the prediction of non-CpG-related promoters. A great deal of prior information is available to focus the search, including the Pol II ChIP-chip or ChIP-seq data, EST or mRNA alignment, and the predicted regions from gene-finding programs. Recent studies suggest that the accuracy of core-promoter prediction can be greatly improved if it is combined with gene prediction (Bajic et al. 2006) or Pol II ChIP signals (Zhao et al. 2007).

We applied CoreBoost_HM to predict the core-promoters of miRNA genes and found that CoreBoost_HM performs well on the six human intergenic miRNA core-promoters which (to the best of our knowledge) are the only known ones. This indicates that CoreBoost_HM can be also used to analyze the promoters of noncoding genes. Many of our predicted miRNA gene promoters show higher PhastCons scores than expected by chance (Supplemental Fig. S2). Further analysis of these putative promoters, e.g., by using comparative genomic approaches, can help to better characterize regulatory regions of miRNA genes (Wang et al. 2008a). Additionally, some intronic miRNA genes appear to have their own promoters; this suggests that the transcriptional regulation of intronic miRNA genes may be more complicated than expected. Further integration of such promoter predictions with other information, such as Pol II binding signal, CAGE tags, ESTs, and genome conservation (Saini et al. 2007), will undoubtedly help to better annotate all miRNA primary transcripts in the genome.

Methods

Genomic sequences and gene annotation

The human genomic sequences (hg18, NCBI build 36) and KnownGene annotation were downloaded from UCSC Genome Browser (http://genome.ucsc.edu/). miRNA sequences and their genomics coordinates were downloaded from miRBase (release 10) (Griffiths-Jones et al. 2006). The CAGE tags were extracted from RIKEN (Carninci et al. 2006). Adjacent CAGE tags within 25 bp on the same DNA strand are clustered, and only CAGE tag clusters with at least three tags were considered. With this criterion, we obtained 89,124 CAGE tag clusters in total. The promoters of the 12,000 genes analyzed by Barski et al. (2007) with known expression information in CD4+ T-cell were used in this analysis. For each of these genes, we collected its TSS from EPD (Cavin Perier et al. 1998) and DBTSS (Wakaguri et al. 2008) database. We defined a promoter to be CpG-related if there is a CpG island locates within its upstream 2 kb to downstream 500-bp region (Xuan et al. 2005) according to UCSC Genome Browser's CpG island annotations. To construct a high-quality training set, we excluded the promoters that contain multiple TSS within 1.5 kb or the annotated TSS are more than 500 bp apart from the 5′-end the annotated genes. After removing redundancy, we got 4263 CpG-related and 1683 non-CpG-related promoters, respectively, which composed our training set. For each promoter in the training set, we chose the annotated TSS as the positive sample and randomly selected one sample from the upstream [−1200, −300] and one from the downstream [300, 1200] regions as negative samples to train the boosting classifier. To further test the performance of the program, we also collected all the 1642 nonoverlapping promoters regions (from −5 kb to 5 kb relative to the gene 5′-end) with at least one TSS annotation from the rest of the genes as our test set. After combining the TSS less than 100 bp apart, this test set contains 2619 independent core-promoters according to EPD and DBTSS annotation. It should be clear that, both the training and test sets are constructed without considering gene expression level, so both active and repressed promoters for a given cell type are included.

Histone modification ChIP-seq data

The histone modification data set was published by Barski et al. (2007) and Wang et al. (2008b). It includes the genome-wide high-resolution profiling of 20 histone methylations, 18 acetylations, one histone variant H2A.Z, and also Pol II binding in human CD4+ T-cells. They detected the number of reads for each histone marker by directly sequencing ChIP pull-down DNA samples using the Illumina 1G Genome Analyzer (ChIP-seq). For data set of each histone marker, we calculated the intensity profile along the genome at a 25-bp resolution. For each position, we counted the number of uniquely aligned reads that are oriented toward it within 100 bp flanking, which is half of the length of DNA fragments in each library (~200 bp). The Pol II signal in CD4+ T-cell was derived from the same ChIP-seq data set. For each position, we counted the number of uniquely aligned reads that are oriented toward it within 35 bp flanking, which is about half of the length of DNA fragments in the library. The positions with more than four reads counts were regarded as Pol II–positive regions (*P*-value $< 1 \times 10^{-6}$ according to background Poisson distribution), and any two Pol II–positive positions closer than 100 bp were clustered. In total, we got 35,772 Pol II–positive regions.

Feature extraction

The features used by CoreBoost_HM can be grouped in two major categories: histone modification profiles and DNA sequence features. The DNA sequence features were extracted using the same method described in CoreBoost (Zhao et al. 2007), including motif features of core-promoter elements and TFBSs, teranucleotide flexibility and energy properties of promoter DNA, sequence features from Markovian modeling of promoter sequences, and *k*mer frequencies (for more details, see Zhao et al. 2007). For each histone modification signal, the 25-bp-resolution intensity profile in a 1.2-kb window was used to construct the vector of local profile of the middle position in this window, and a vector of empirical average pattern was calculated by averaging the profile of this modification marker on all training promoter regions centered by TSS. Larger window sizes were also tested in our experiments, but no significant improvement was observed in prediction accuracy. We then encoded each histone modification signal with two features: one is the Pearson correlation coefficient between the vector of local profile and the vector of empirical average pattern of this modification signal around the TSS, which is used to capture the shape similarity; the other is the dot product of these two vectors, which reflects the weighted intensity of this modification signal. The raw histone feature table includes 78 features for the 20 methylations, 18 acetylations, and one histone variant H2A.Z.

LogitBoost with stumps

Boosting is a supervised machine learning algorithm combining many weak classifiers to create a single strong classifier (Hastie et al. 2001). In this work, we chose stumps as the weak classifiers, which are single-split decision trees with only two terminal nodes (Breiman et al. 1984). Denote the training data as $(x_1, y_1), \ldots, (x_N, y_N)$, where x_i is the feature vector and y_i is the class label $\{-1, 1\}$. We defined $f_m(x)$ as the *m*th weak binary classifier producing value of +1 or −1, and $F(x)=\sum_{m=1}^{M} c_m f_m(x)$ as the ensemble of a series of *M* weak classifiers, where c_m are constants and *M* is determined by cross-validation. Let $y^* = (y + 1)/2$, taking values from {0, 1}. We implemented LogitBoost algorithm (Friedman et al. 2000) to directly estimate the posterior class probability:

$$P(Y = 1|X = x) = \frac{\exp(F(x))}{\exp(F(x)) + \exp(-F(x))}$$

(for more details, see Zhao et al. 2007).

The output prediction score of CoreBoost_HM is not identical as CoreBoost. In CoreBoost, we combined the output of two binary classifiers that were designed to classify the promoter against the upstream sequence and downstream sequence, respectively. In CoreBoost_HM, we used negative samples selected from random positions in upstream and downstream of TSS and only implemented one classifier against both the upstream and downstream sequence. Additionally, we noticed that the true promoters and their proximal adjacent regions generally have higher average score than the non-promoter regions, and the false-positive predictions typically only have a single narrow peak, so we smoothed the output score to filter the noises and improve the performance. We defined the raw CoreBoost_HM score as

$$\begin{aligned} CoreBoost\_HM\_Score &= -\ln(1 - P(Y = 1|X = x)) \\ &= \ln(1 + \exp(2^*F(x))), \end{aligned}$$

which is then smoothed with an optimized window of size 100 bp (Supplemental Fig. S3).

Performance evaluation

We ran the McPromoter and CoreBoost with default settings and got the raw prediction score profiles for every 10 bp. As EP3 gives the best performance when setting the window size to 400 bp (at this window size, EP3 gives a prediction value for every 400 bp) (Abeel et al. 2008) to get higher resolution predictions, according to the recommendation of the investigators of EP3, we included 10 bp each time at the beginning of the sequence and ran the program 40 times to get the prediction values for every 10 bp. All the comparisons among predictors were based on these raw score profiles.

Sensitivity, positive predictive value (PPV) (Bajic et al. 2004) and *F*-score (Abeel et al. 2008) were used to evaluate the performance of the different predictors, which are defined as

$$Sensitivity = \frac{TP}{TP + FN},$$

$$PPV = \frac{TP}{TP + FP},$$

$$F = \frac{2}{(1/Sensitivity) + (1/PPV)} = \frac{2{*}Sensitivity{*}PPV}{Sensitivity + PPV},$$

where TP is the true positives, TN is the true negatives, FP is the false positives, and FN is the false negatives. *F*-score is the harmonic average of sensitivity and PPV. For each program, we directly counted these numbers from their raw prediction score profile. If any two predictions above a certain threshold were less than 500 bp apart, these two predictions were clustered and only the position with a better score was kept as the representative. We did this clustering iteratively, and only the peak of each cluster was counted as predicted TSS. When we evaluated the performance of each program, we compared the relative distance of the peak position to the annotated TSS. A prediction was counted as a true positive only if the relative distance is smaller than a given distance resolution (50, 200, and 500 bp, respectively).

Availability

The CoreBoost_HM web interface is available at http://rulai.cshl.edu/tools/CoreBoost_HM. The training and testing data sets used in this analysis and the CoreBoost_HM predictions on miRNAs can be downloaded from this website.

Acknowledgments

We thank Dr. Andrew Smith and Dr. Chaolin Zhang for helpful discussion. This work was supported by NIH grant HG001696 and NSFC grants 60572086, 60775002, and 60805009.

References

Abeel, T., Saeys, Y., Bonnet, E., Rouze, P., and Van de Peer, Y. 2008. Generic eukaryotic core promoter prediction using structural features of DNA. *Genome Res.* **18:** 310–323.

Bajic, V.B., Seah, S.H., Chong, A., Zhang, G., Koh, J.L., and Brusic, V. 2002. Dragon Promoter Finder: Recognition of vertebrate RNA polymerase II promoters. *Bioinformatics* **18:** 198–199.

Bajic, V.B., Tan, S.L., Suzuki, Y., and Sugano, S. 2004. Promoter prediction analysis on the whole human genome. *Nat. Biotechnol.* **22:** 1467–1473.

Bajic, V.B., Brent, M.R., Brown, R.H., Frankish, A., Harrow, J., Ohler, U., Solovyev, V.V., and Tan, S.L. 2006. Performance assessment of promoter predictions on ENCODE regions in the EGASP experiment. *Genome Biol.* **7:** 1–13.

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129:** 823–837.

Bartel, D.P. 2004. MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell* **116:** 281–297.

Borchert, G.M., Lanier, W., and Davidson, B.L. 2006. RNA polymerase III transcribes human microRNAs. *Nat. Struct. Mol. Biol.* **13:** 1097–1101.

Breiman, L., Friedman, J., Olshen, R., and Storne, C. 1984. *Classification and regression trees*. Wadsworth International Group, Belmont, CA.

Cai, X., Hagedorn, C.H., and Cullen, B.R. 2004. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA* **10:** 1957–1966.

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.

Cavin Perier, R., Junier, T., and Bucher, P. 1998. The Eukaryotic Promoter Database EPD. *Nucleic Acids Res.* **26:** 353–357.

Chang, T.C., Wentzel, E.A., Kent, O.A., Ramachandran, K., Mullendore, M., Lee, K.H., Feldmann, G., Yamakuchi, M., Ferlito, M., Lowenstein, C.J., et al. 2007. Transactivation of miR-34a by p53 broadly influences gene expression and promotes apoptosis. *Mol. Cell* **26:** 745–752.

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.

Friedman, J., Hastie, T., and Tibshirani, R. 2000. Additive logistic regression: A statistical view of boosting. *Ann. Stat.* **28:** 337–407.

Fukao, T., Fukuda, Y., Kiga, K., Sharif, J., Hino, K., Enomoto, Y., Kawamura, A., Nakamura, K., Takeuchi, T., and Tanabe, M. 2007. An evolutionarily conserved mechanism for microRNA-223 expression revealed by microRNA gene profiling. *Cell* **129:** 617–631.

Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A., and Enright, A.J. 2006. miRBase: MicroRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.* **34:** D140–D144.

Hastie, T., Tibshirani, R., and Friedman, J. 2001. *The elements of statistical learning: Data mining, inference, and prediction*. Springer-Verlag, New York.

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A., et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39:** 311–318.

Houbaviy, H.B., Dennis, L., Jaenisch, R., and Sharp, P.A. 2005. Characterization of a highly variable eutherian microRNA gene. *RNA* **11:** 1245–1257.

Kouzarides, T. 2007. Chromatin modifications and their function. *Cell* **128:** 693–705.

Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A.O., Landthaler, M., et al. 2007. A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* **129:** 1401–1414.

Lee, Y., Kim, M., Han, J., Yeom, K.H., Lee, S., Baek, S.H., and Kim, V.N. 2004. MicroRNA genes are transcribed by RNA polymerase II. *EMBO J.* **23:** 4051–4060.

Marson, A., Levine, S.S., Cole, M.F., Frampton, G.M., Brambrink, T., Johnstone, S., Guenther, M.G., Johnston, W.K., Wernig, M., Newman, J., et al. 2008. Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* **134:** 521–533.

Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P., et al. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448:** 553–560.

Ohler, U., Niemann, H., Liao, G., and Rubin, G.M. 2001. Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition. *Bioinformatics* **17:** S199–S206.

Packer, M.J., Dauncey, M.P., and Hunter, C.A. 2000. Sequence-dependent DNA structure: Tetranucleotide conformational maps. *J. Mol. Biol.* **295:** 85–103.

Rodriguez, A., Griffiths-Jones, S., Ashurst, J.L., and Bradley, A. 2004. Identification of mammalian microRNA host genes and transcription units. *Genome Res.* **14:** 1902–1910.

Roh, T.Y., Cuddapah, S., Cui, K., and Zhao, K. 2006. The genomic landscape of histone modifications in human T cells. *Proc. Natl. Acad. Sci.* **103:** 15782–15787.

Saini, H.K., Griffiths-Jones, S., and Enright, A.J. 2007. Genomic analysis of human microRNA transcripts. *Proc. Natl. Acad. Sci.* **104:** 17719–17724.

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.* **103:** 1412–1417.

Schones, D.E. and Zhao, K. 2008. Genome-wide approaches to studying chromatin modifications. *Nat. Rev. Genet.* **9:** 179–191.

Schones, D.E., Cui, K., Cuddapah, S., Roh, T.Y., Barski, A., Wang, Z., Wei, G., and Zhao, K. 2008. Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132:** 887–898.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15:** 1034–1050.

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.

Wakaguri, H., Yamashita, R., Suzuki, Y., Sugano, S., and Nakai, K. 2008. DBTSS: Database of transcription start sites, progress report 2008. *Nucleic Acids Res.* **36:** D97–D101.

Wang, X., Gu, J., Zhang, M.Q., and Li, Y. 2008a. Identification of phylogenetically conserved microRNA *cis*-regulatory elements across 12 *Drosophila* species. *Bioinformatics* **24:** 165–171.

Wang, Z., Zang, C., Rosenfeld, J.A., Schones, D.E., Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Peng, W., Zhang, M.Q., et al. 2008b. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.* **40:** 897–903.

Woods, K., Thomson, J.M., and Hammond, S.M. 2007. Direct regulation of an oncogenic micro-RNA cluster by E2F transcription factors. *J. Biol. Chem.* **282:** 2130–2134.

Xuan, Z., Zhao, F., Wang, J., Chen, G., and Zhang, M.Q. 2005. Genome-wide promoter extraction and analysis in human, mouse, and rat. *Genome Biol.* **6:** R72. doi: 10.1186/gb-2005-6-8-r72.

Zhang, M.Q. 1998. Identification of human gene core promoters in silico. *Genome Res.* **8:** 319–326.

Zhang, M.Q. 2007. Computational analyses of eukaryotic promoters. *BMC Bioinformatics* **8:** S3. doi: 10.1186/1471-2105-8-S6-S3.

Zhao, X., Xuan, Z., and Zhang, M.Q. 2007. Boosting with stumps for predicting transcription start sites. *Genome Biol.* **8:** R17. doi: 10.1186/gb-2007-8-2-r17.

Zhou, X., Ruan, J., Wang, G., and Zhang, W. 2007. Characterization and identification of microRNA core promoters in four model species. *PLoS Comput. Biol.* **3:** e37. doi: 10.1371/journal.pcbi.0030037.

Received June 3, 2008; accepted in revised form October 27, 2008.

High-resolution human core-promoter prediction with CoreBoost_HM

Xiaowo Wang, Zhenyu Xuan, Xiaoyue Zhao, et al.

Genome Res. 2009 19: 266-275 originally published online November 7, 2008
Access the most recent version at doi:10.1101/gr.081638.108