

Supplementary figures and Tables

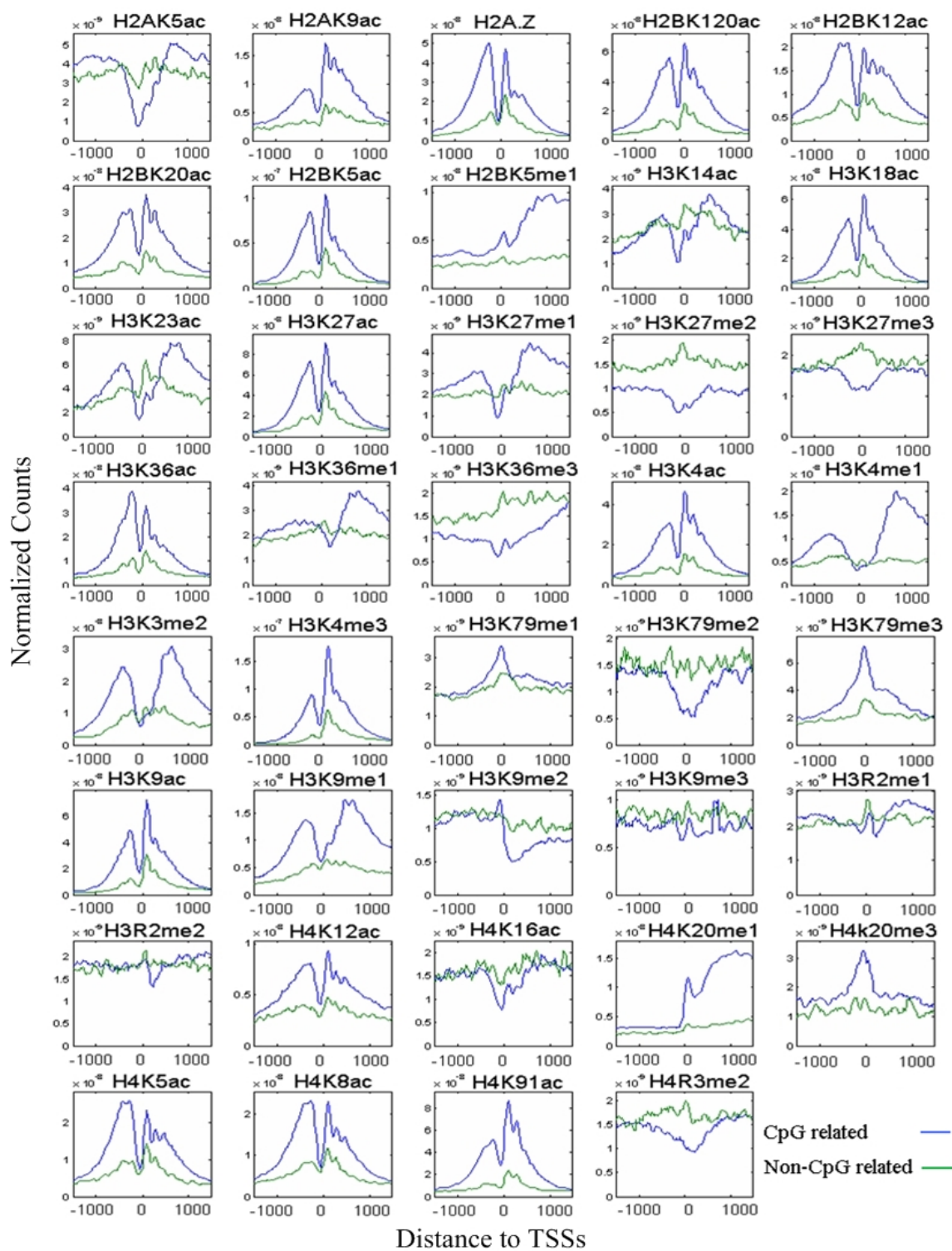


Figure S1. Histone modification profiles around gene transcription start sites (TSSs) of all 39 histone markers used in this study.

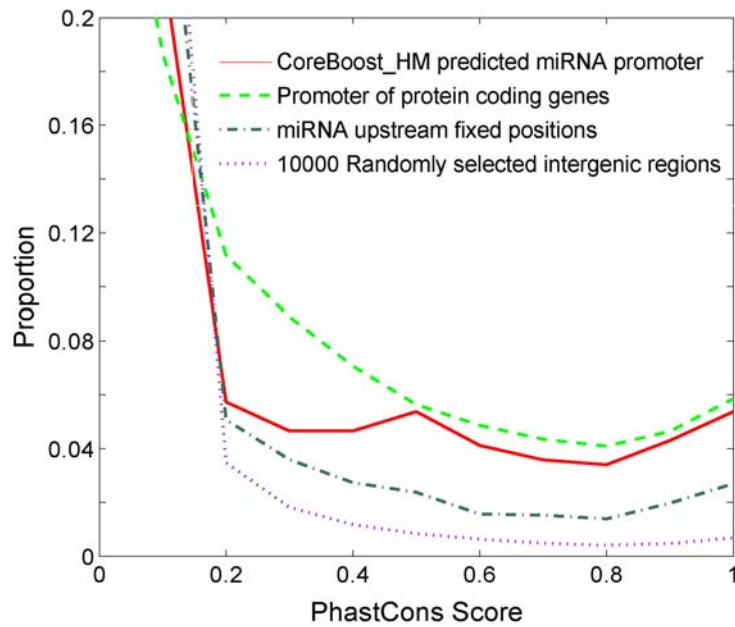


Figure S2. CoreBoost_HM predicted core-promoters of intergenic miRNA have significantly higher PhastCons Score than expected by chance.

This figure shows the histogram of the PhastCons Score of CoreBoost_HM predicted miRNA promoters (with higher specificity parameter settings, see Supplementary S12), promoter of protein coding genes, fixed upstream positions (upstream 1 kb and 2 kb) of intergenic miRNAs, and 10000 randomly selected intergenic positions. For each promoter (or genomic position), we calculated the average PhastCons score (based on the alignments of 16 vertebrates with human) of the 100bp region centred at that position.

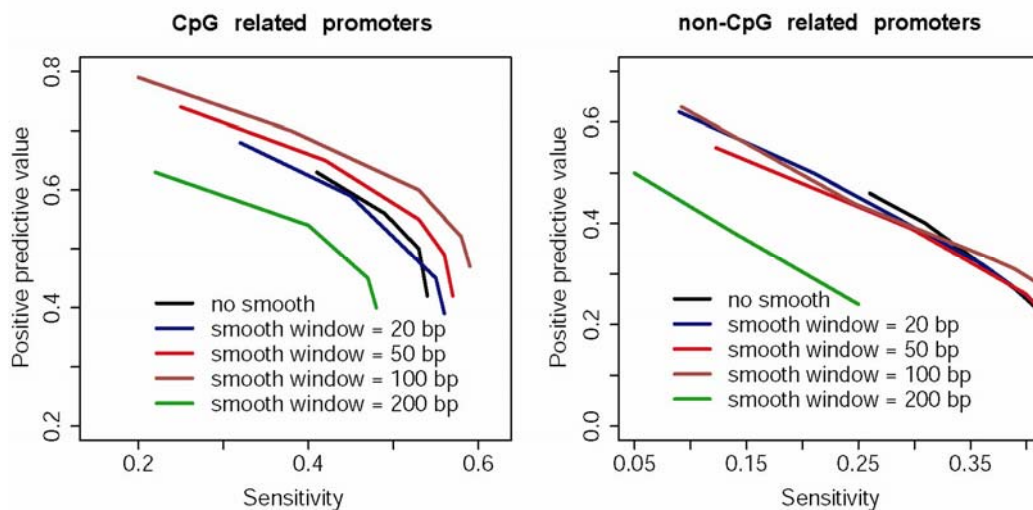


Figure S3. Smooth window size selection for CoreBoost_HM prediction results

This figure shows the plot of positive predictive values (PPV) versus sensitivity for CpG (A) and non-CpG (B) related promoters when using smooth window of different size.

Only a prediction locating within 50 bp of the annotated TSS is regarded as a true positive hit and the prediction results are clustered within 500 bp. This figure shows that smoothing the output score to filter the noises can improve the prediction accuracy (especially for the CpG related predictor), and the window size of 100 bp can achieve the best performance.

Supplementary Tables

All the performance evaluation results shown in Table S2, S3, S5 and S7 were estimated by searching the 2.4 kb region centered at the TSS of each promoter in the training set, and ten-fold cross-validation was performed for CoreBoost_HM and the boosting classifier with only histone modification features (BHMF). We varied different thresholds of prediction score for each predictor to choose the threshold with maximum F score in each condition.

Table S1. Histone features contributing to the BHMF classifier (Excel file)

Table S2. non-CpG related core-promoter prediction using histone modification markers (on training set)

Type	Number of Histone markers ^a	Maximum allowed distance from true TSSs								
		50			200			500		
		Sen	PPV	F	Sen	PPV	F	Sen	PPV	F
Methylation	Top one	0.07	0.09	0.08	0.36	0.30	0.32	0.65	0.47	0.55
	Top three	0.11	0.16	0.13	0.35	0.33	0.34	0.75	0.46	0.57
	Top five	0.12	0.19	0.15	0.37	0.32	0.35	0.80	0.45	0.58
	Top ten	0.14	0.17	0.15	0.39	0.32	0.35	0.75	0.47	0.58
	All 20 Markers	0.11	0.22	0.15	0.35	0.35	0.35	0.84	0.45	0.58
Acetylation	Top one	0.07	0.08	0.08	0.37	0.22	0.28	0.62	0.45	0.52
	Top three	0.07	0.09	0.08	0.44	0.21	0.29	0.80	0.42	0.55
	Top five	0.11	0.08	0.10	0.40	0.24	0.30	0.73	0.45	0.56
	Top ten	0.10	0.11	0.11	0.39	0.26	0.31	0.67	0.47	0.57
	All 18 Markers	0.10	0.11	0.11	0.39	0.25	0.31	0.71	0.47	0.57
All 39 Histone Markers	Top three	0.08	0.13	0.10	0.33	0.35	0.34	0.76	0.46	0.57
	Top five	0.10	0.18	0.13	0.37	0.33	0.35	0.82	0.45	0.58
	Top ten	0.15	0.17	0.16	0.39	0.33	0.36	0.64	0.53	0.58
	All 39 Markers	0.13	0.21	0.16	0.37	0.34	0.36	0.71	0.50	0.59

^aThe corresponding top histone markers are listed in Table 1 of main text.

Table S3. CpG related core-promoter prediction using histone modification markers (on training set)

Type	Number of Histone markers ^a	Maximum allowed distance from true TSSs								
		50 bp			200 bp			500 bp		
		Sen	PPV	F	Sen	PPV	F	Sen	PPV	F
Methylation	Top one	0.12	0.12	0.12	0.62	0.30	0.40	0.92	0.59	0.72
	Top three	0.15	0.28	0.21	0.52	0.55	0.54	0.83	0.73	0.77
	Top five	0.22	0.34	0.27	0.59	0.59	0.59	0.79	0.81	0.80
	Top ten	0.26	0.41	0.32	0.61	0.65	0.63	0.84	0.81	0.82
	All 20 Markers	0.27	0.49	0.35	0.60	0.71	0.65	0.83	0.85	0.84
Acetylation	Top one	0.13	0.11	0.12	0.49	0.28	0.36	0.76	0.55	0.64
	Top three	0.13	0.22	0.17	0.51	0.40	0.45	0.78	0.67	0.72
	Top five	0.17	0.28	0.21	0.49	0.52	0.50	0.79	0.68	0.73
	Top ten	0.22	0.36	0.27	0.50	0.57	0.54	0.85	0.67	0.75
	All 18 Markers	0.24	0.37	0.30	0.53	0.58	0.56	0.80	0.74	0.77
All 39 Histone Markers	Top three	0.16	0.31	0.22	0.50	0.60	0.54	0.68	0.82	0.74
	Top five	0.21	0.31	0.25	0.55	0.61	0.60	0.85	0.75	0.80
	Top ten	0.26	0.41	0.32	0.61	0.68	0.64	0.84	0.82	0.83
	All 39 Markers	0.33	0.46	0.38	0.68	0.69	0.68	0.87	0.85	0.86

^aThe corresponding top histone markers are listed in Table 1 of main text.

Table S4. Top histone markers contributing to the prediction of active and repressed promoters

Promoter Type	CpG related		Non-CpG related	
Related gene expression level	High	Low	High	Low
Top Histone Markers	H3K4me3	H3K4me3	H3K4me3	H3K4me3
	H3K4me1	H3K4me2	H3K79me3	H2A.Z
	H3K79me3	H3K4me1	H2A.Z	H3K27ac
	H3K4me2	H4K20me1	H4R3me2	H3K18ac
	H3K23ac	H3K79me3	H3K79me1	H2BK120ac
	H2AK5ac	H2AZ	H3K27ac	H3K27me3
	H2AZ	H2AK5ac	H4K20me3	H3K79me3
	H3K27me1	H4K20me3	H3K4me2	H4K16ac
	H3K36me1	H3K9me2	H3K36me3	H4R3me2
	H3K9me2	H4K5ac	H3K9me2	H3K79me1

Table S5. Correlation between the prediction power and gene expression (on training set)

Method	Promoter Type	Related gene expression level	Maximum allowed distance from true TSSs								
			50 bp			200 bp			500 bp		
			Sen	PPV	F	Sen	PPV	F	Sen	PPV	F
BHMF	CpG related	High	0.39	0.54	0.45	0.76	0.76	0.76	0.94	0.88	0.91
		Low	0.26	0.37	0.31	0.60	0.60	0.60	0.85	0.77	0.81
	non-CpG related	High	0.17	0.31	0.22	0.45	0.40	0.42	0.76	0.53	0.63
		Low	0.10	0.10	0.10	0.36	0.25	0.30	0.80	0.44	0.57
CoreBoost_HM	CpG related	High	0.61	0.63	0.62	0.80	0.83	0.82	0.95	0.90	0.93
		Low	0.46	0.57	0.51	0.74	0.68	0.71	0.89	0.85	0.87
	non-CpG related	High	0.41	0.36	0.38	0.62	0.53	0.57	0.76	0.66	0.71
		Low	0.35	0.29	0.32	0.55	0.47	0.51	0.72	0.65	0.68

Table S6. Features contributing to CoreBoost_HM (Excel file)

Table S7. Performance evaluation on training set

Program	CpG related promoter									non-CpG related promoter								
	50			200			500			50			200			500		
	Sen	PPV	F	Sen	PPV	F	Sen	PPV	F	Sen	PPV	F	Sen	PPV	F	Sen	PPV	F
CoreBoost_HM	0.53	0.60	0.57	0.79	0.73	0.76	0.92	0.88	0.90	0.37	0.33	0.35	0.58	0.50	0.54	0.74	0.65	0.69
BHMA	0.33	0.46	0.38	0.68	0.69	0.68	0.87	0.85	0.86	0.13	0.21	0.16	0.37	0.34	0.36	0.71	0.50	0.59
CoreBoost	0.34	0.37	0.35	0.69	0.55	0.61	0.92	0.74	0.82	0.26	0.38	0.31	0.50	0.44	0.47	0.65	0.58	0.61
McPromoter	0.26	0.24	0.25	0.53	0.48	0.51	0.91	0.65	0.75	0.23	0.18	0.20	0.44	0.35	0.39	0.75	0.49	0.59
EP3	0.19	0.20	0.20	0.69	0.61	0.65	0.93	0.85	0.89	0.15	0.12	0.13	0.56	0.36	0.44	0.81	0.54	0.65

Table S8. Performance evaluation on test set based on the combined annotation of DBTSS, EPD and CAGE tag clusters.

Program	Maximum allowed distance from true TSSs								
	50 bp			200 bp			500 bp		
	Sen ^a	PPV ^b	F ^c	Sen	PPV	F	Sen	PPV	F
CoreBoost_HM	0.24	0.43	0.31	0.42	0.66	0.51	0.60	0.81	0.69
BHMF	0.15	0.39	0.22	0.33	0.69	0.45	0.52	0.83	0.64
CoreBoost	0.19	0.26	0.22	0.39	0.46	0.43	0.62	0.60	0.61
McPromoter	0.22	0.20	0.21	0.41	0.46	0.43	0.66	0.67	0.67
EP3	0.14	0.22	0.17	0.41	0.55	0.47	0.63	0.75	0.69

Table S9. False positive rate of CoreBoost_HM estimated by searching 1000 randomly selected 10kb long intronic regions which do not overlap with any DBTSS annotation, CAGE tag clusters or Pol II peaks.

Threshold	Estimated false positive number ^b per 10 kb					
	CpG related predictor			Non-CpG related predictor		
	cd ^a 500 bp	cd 1000 bp	cd 2000 bp	cd 500 bp	cd 1000 bp	cd 2000 bp
0.5	0.743	0.688	0.632	9.080	5.775	3.273
0.6	0.254	0.237	0.229	6.217	4.781	3.021
0.7	0.113	0.108	0.107	3.613	3.224	2.421
0.8	0.046	0.045	0.044	1.831	1.762	1.522
0.9	0.015	0.015	0.015	0.857	0.850	0.799
1.0	0.009	0.009	0.009	0.331	0.330	0.329
1.1	0.004	0.004	0.004	0.118	0.118	0.118
1.2	0.002	0.002	0.002	0.042	0.042	0.042
1.5	0.000	0.000	0.000	0.001	0.001	0.001

a: cd: clustering distance. If any two predictions above a certain threshold were less than that distance (e.g. 500 bp) apart, these two predictions were clustered and only the position with better score was kept as the representative. We did this clustering iteratively and only the peak of each cluster was counted as predicted TSSs.

b: It should be noticed that as some unknown promoters may locate within these “negative control” regions, the false positive number estimated by this way is higher than the reality.

Table S10. Prediction on known miRNA promoters.

miRNA ID	Relative distance to the 5' end of pre-miRNA				
	Experimentally characterized TSS	CoreBoost_HM	EP3	CoVote ^a	Marson et al. Cell, 2008
mir-23a cluster	-15 bp	-387 bp	-360 bp	-196 bp	128 bp
mir-21	-2445 bp	-2940 bp	-2090 bp	-275 bp	None
mir-17 cluster	-2922 bp	-2955 bp	-1790 bp	None	-2786 bp
mir-34a	-30522 to -30615 bp	-30557 bp	-30620 bp	None	-30561 bp
mir-223	-3409 bp	-3407 bp	-3410 bp	-275 bp	None
mir-371 cluster	-697 bp	364 bp	590 bp	-624 bp	412 bp

For the predictions of Marson et al. 2008 and CoVote, as they only provide a putative promoter region (~1kb for Marson 2008 and ~350 bp for CoVote), we use the distance from the midpoint of their predicted promoter to the true TSS to denote their prediction accuracy.

a. CoVote dose not provide excusable program or source code. It only provides the predictions within upstream 2kb of the pre-miRNAs.

Table S11. CoreBoost_HM predictions on intergenic miRNAs are partially supported by Pol II signal or CAGE tags.

Predicted promoter sets	Number of predictions	Supported by CAGE ^a	Supported by Pol II ^b	CAGE +Pol II
Higher Specificity ^c	145	51	34	65
Mid_sensitivity_specificity	293	69	41	87
Higher Sensitivity	518	82	50	107

a: We stated a predicted core-promoter is supported by CAGE only if it is located within 500bp of a CAGE clusters on same strand.

b: We stated a predicted core-promoter is supported by Pol II signal only if it is located within 500bp of a Pol II positive region.

c: Estimated sensitivity and false positive number are listed in Supplementary Table S8.

Table S12. CoreBoost_HM predictions on upstream 10 kb of intergenic miRNAs

(Excel file)

Table S13. CoreBoost_HM predictions on upstream 10 kb of intronic miRNAs (only the predictions locate within the same intron are listed) (Excel file)