Supplementary Material

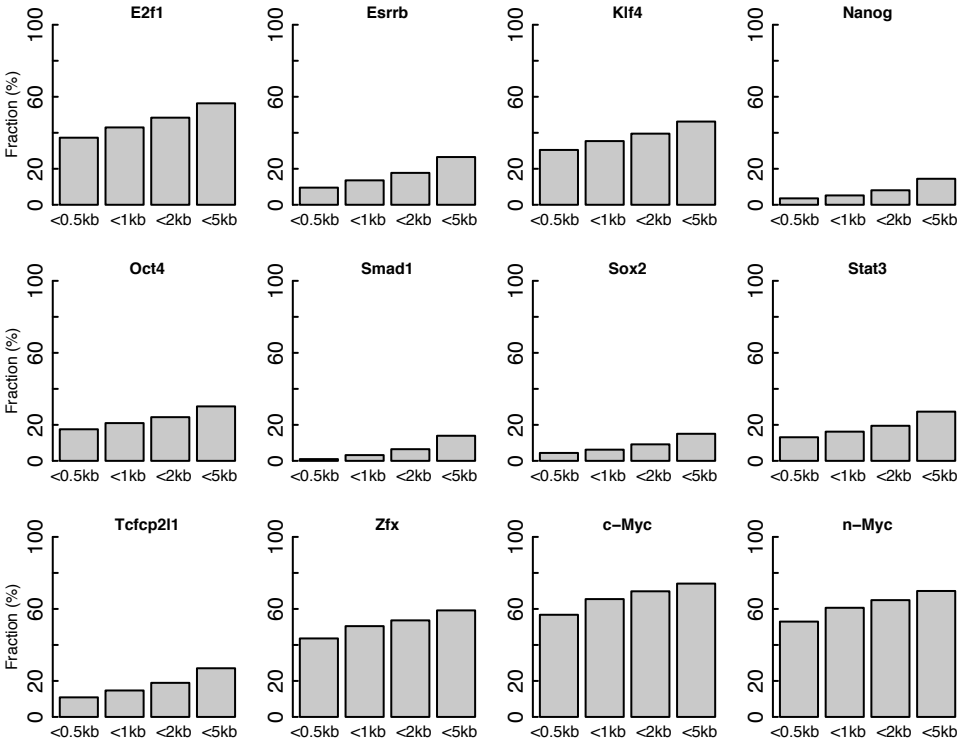

Figure S1: Fraction of transcription factor binding peaks that is nearby transcription start site (TSS) of genes. For each of TF, the percentage of binding peaks within 0.5, 1, 2 and 5kb of the TSS are shown.

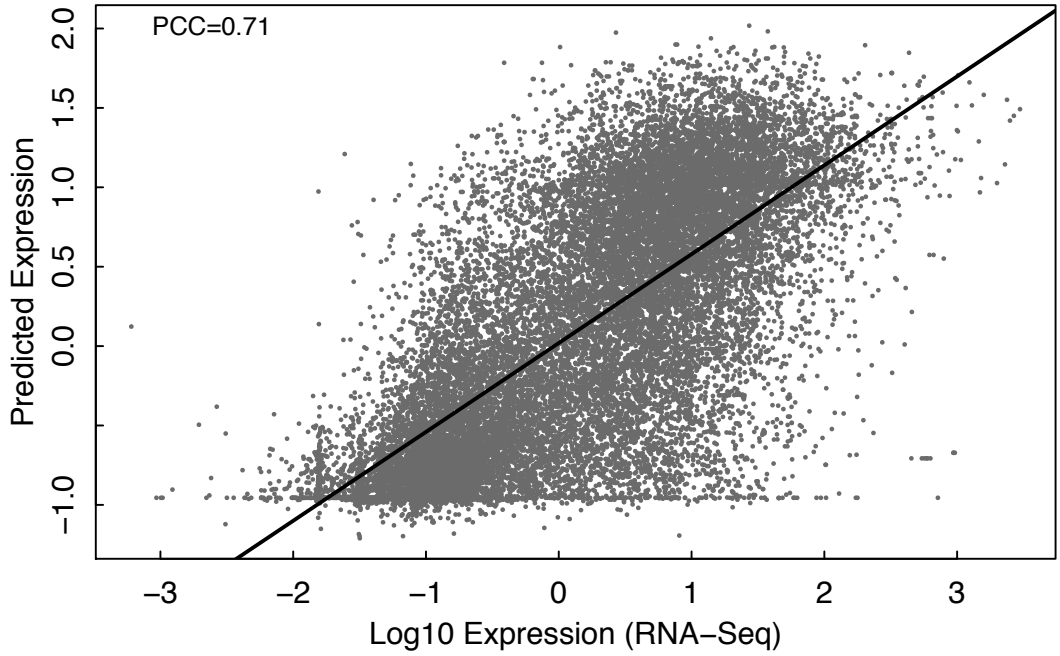

Figure S2: Accuracy of the TF model for gene expression prediction in ESC using binding signals at the TSS. This model achieves the highest prediction accuracy among all bins.

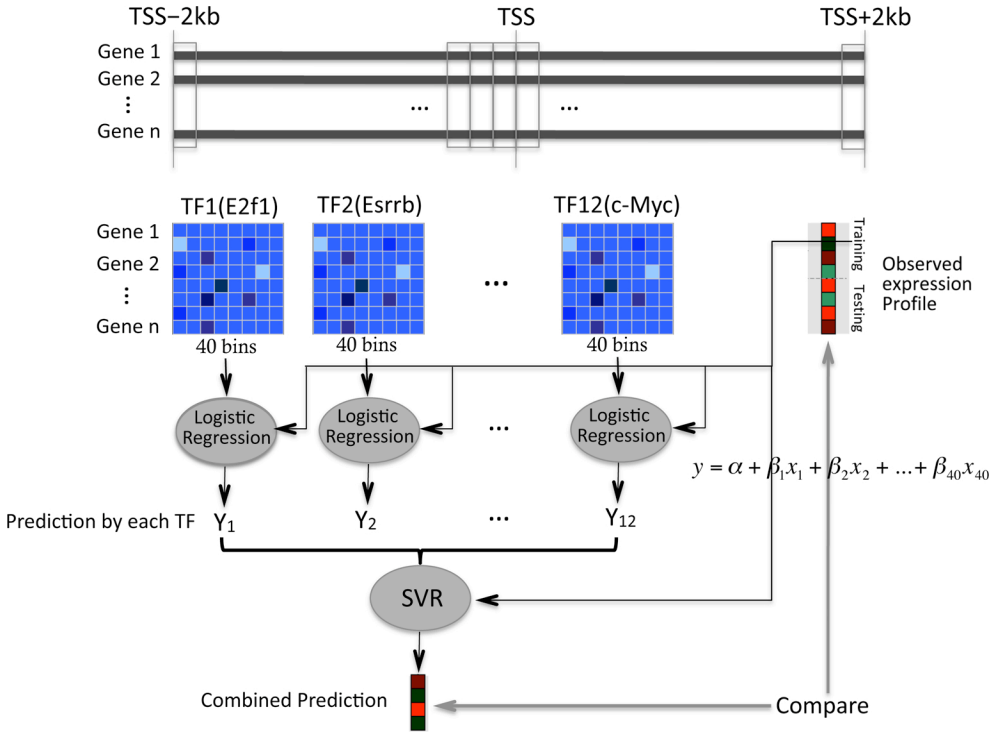

Figure S3: Schematic diagram of the alternative two-layer model for gene expression prediction in ESC. In this model, the first layer predicts gene expression levels based on the binding profile of each transcription factor across 40 bins (from -2kb upstream to 2kb downstream of TSS) using logistic regression. Then in the second layer, the 12 independent predictions by each TF were taken as predictors by the support vector regression algorithm to obtain the final prediction. Prediction accuracy of the model was evaluated by comparing with experimental results.

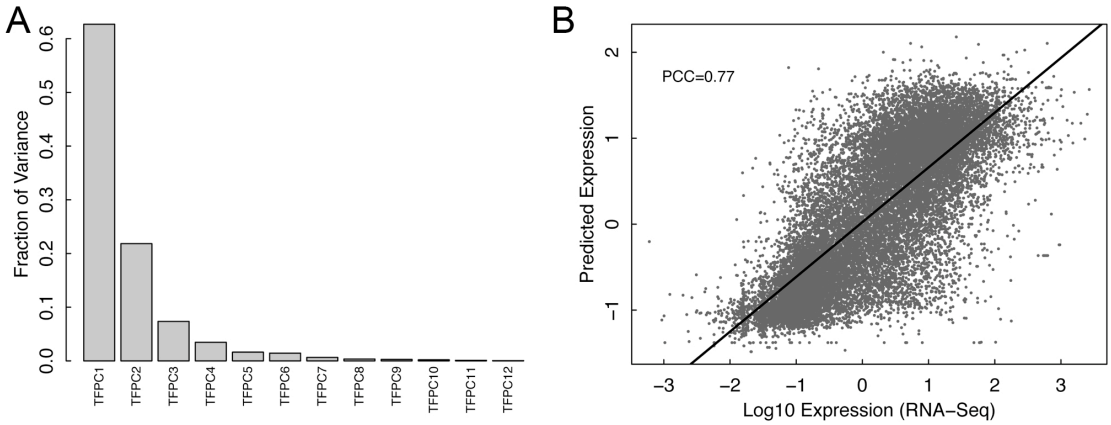

Figure S4: Integrating principle component analysis with SVR model. (A) The fraction of variation of gene expression explained by each TF principle component vector (TFPC). (B) Prediction accuracy of the SVR model using TFPCs as the predictors. The maximum signal matrix (gene × TF) is used for PCA analysis.

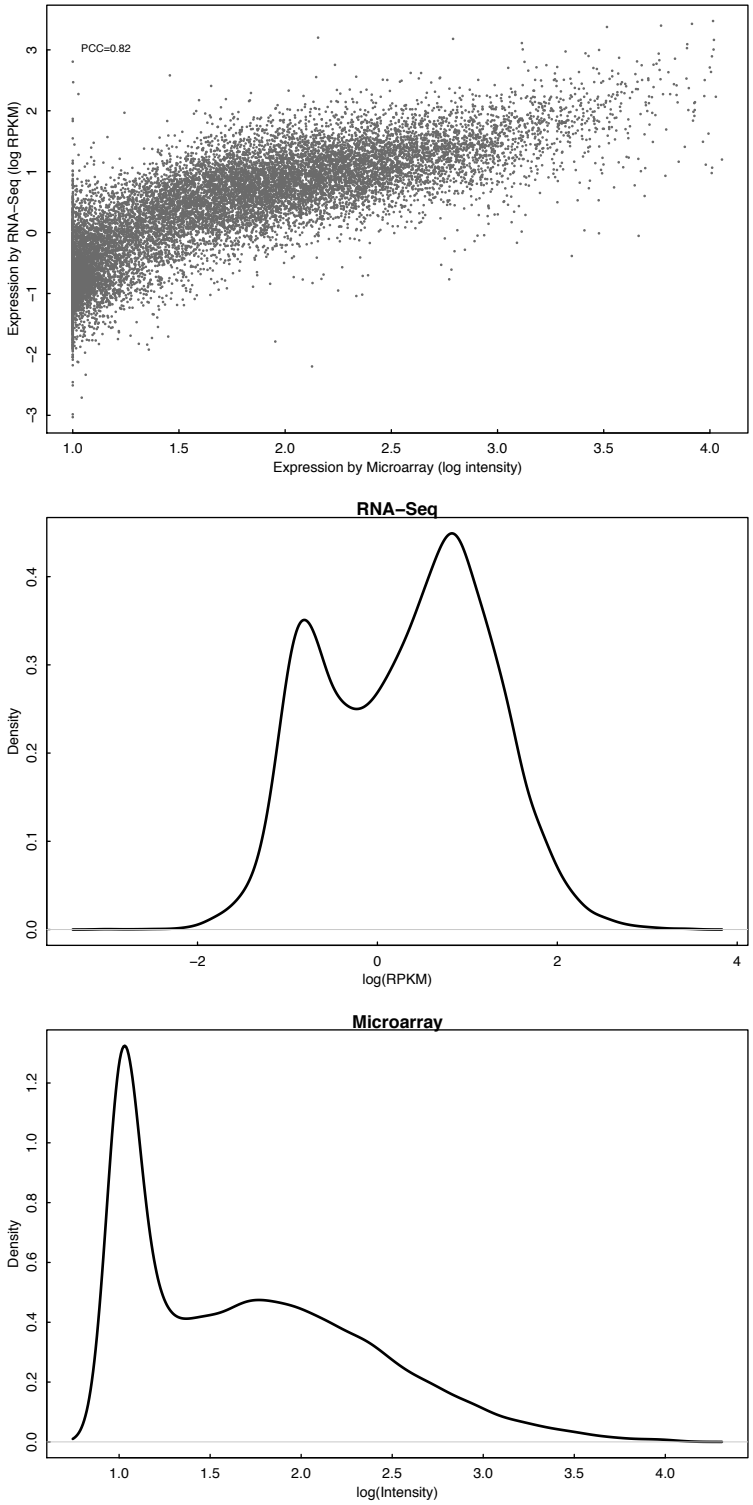

Figure S5: Consistency of gene expression levels measured by microarrays and RNA-Seq experiments. The top figure shows the scatterplot of the measurements by the two methods. The Middle figure shows the distribution of expression levels (log(RPKM)) quantified by RNA-Seq. The bottom figure shows the distribution of expression levels quantified by microarrays.

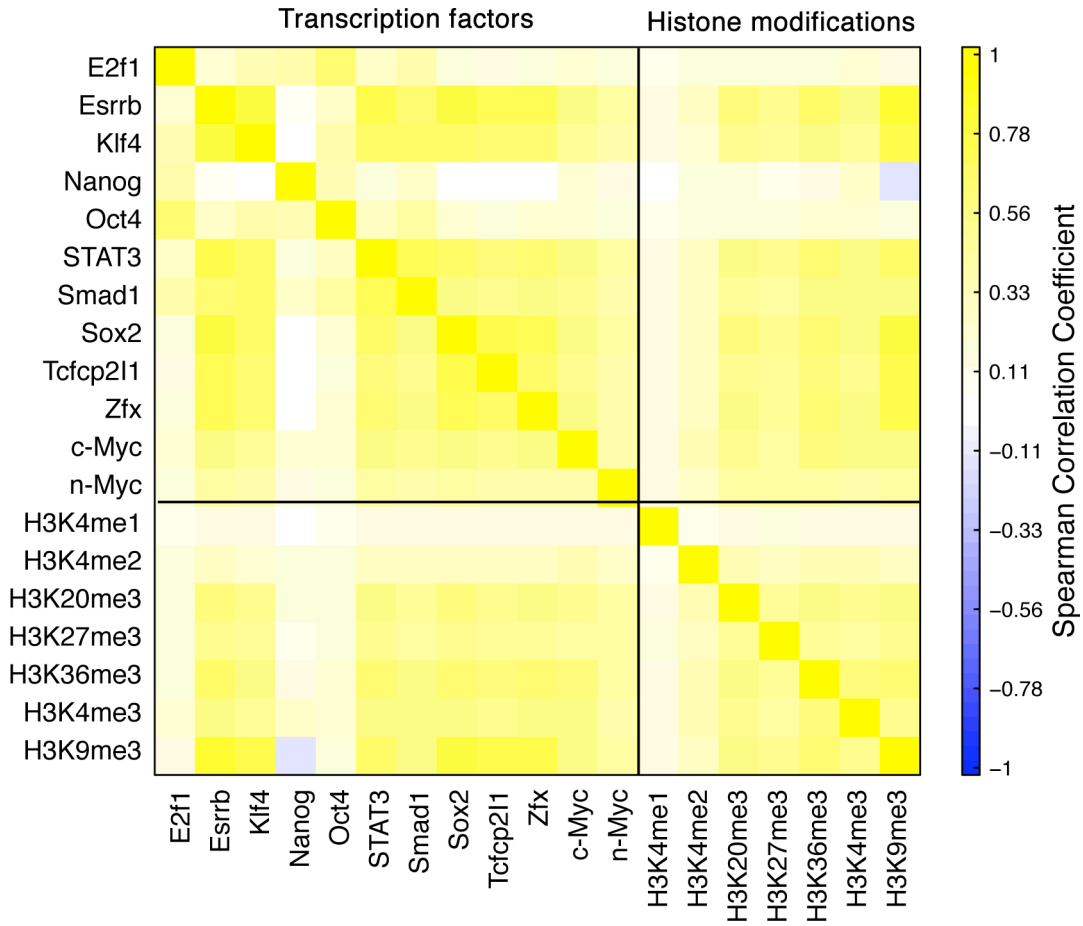

Figure S6: The Correlation of signals between TF binding and histone modifications. The signals surrounding the TSS of each gene (-1kb upstream to 1kb downstream) were averaged to represent the binding strength of a TF or magnitude of a histone modification of the gene.

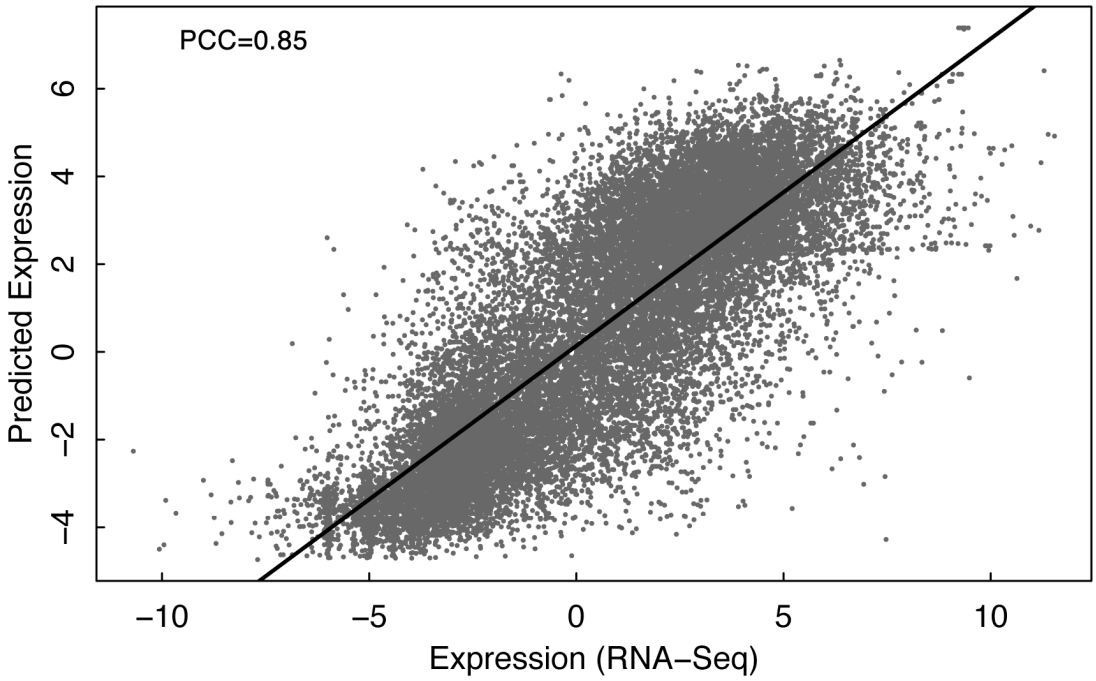

Figure S7: Accuracy of the TF+HM two-layer model for gene expression prediction in ESC. The model integrates the signals from 12 TFs and 7 histone modifications.

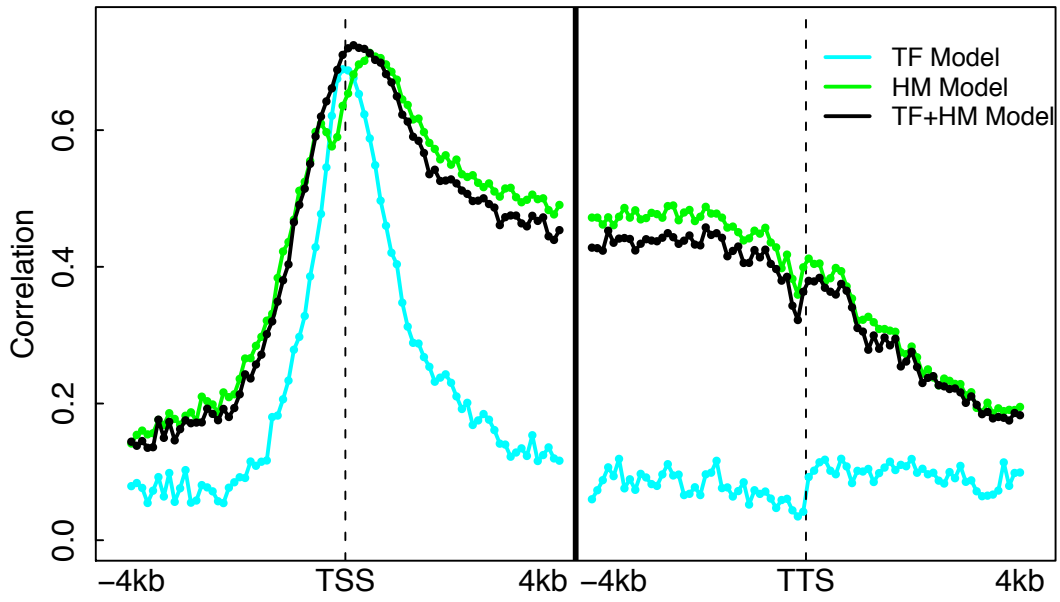

Figure S8: The accuracy of three models for predicting expression levels of the non-overlapping RefSeq genes in 160 bins.

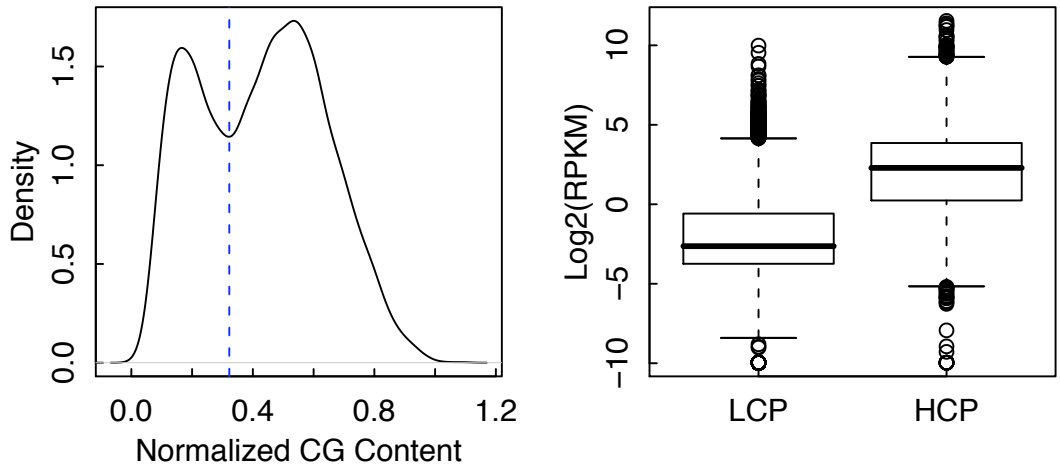

Figure S9: CpG content distinguishes two promoter groups: low CpG content promoters and high CpG content promoters. (A) Distribution of normalized CpG content for mouse RefSeq genes. The DNA regions [-1.5kb, 1.5kb] around the TSS were used for calculating normalized CpG content. (B) The expression levels of the LCP and HCP genes.

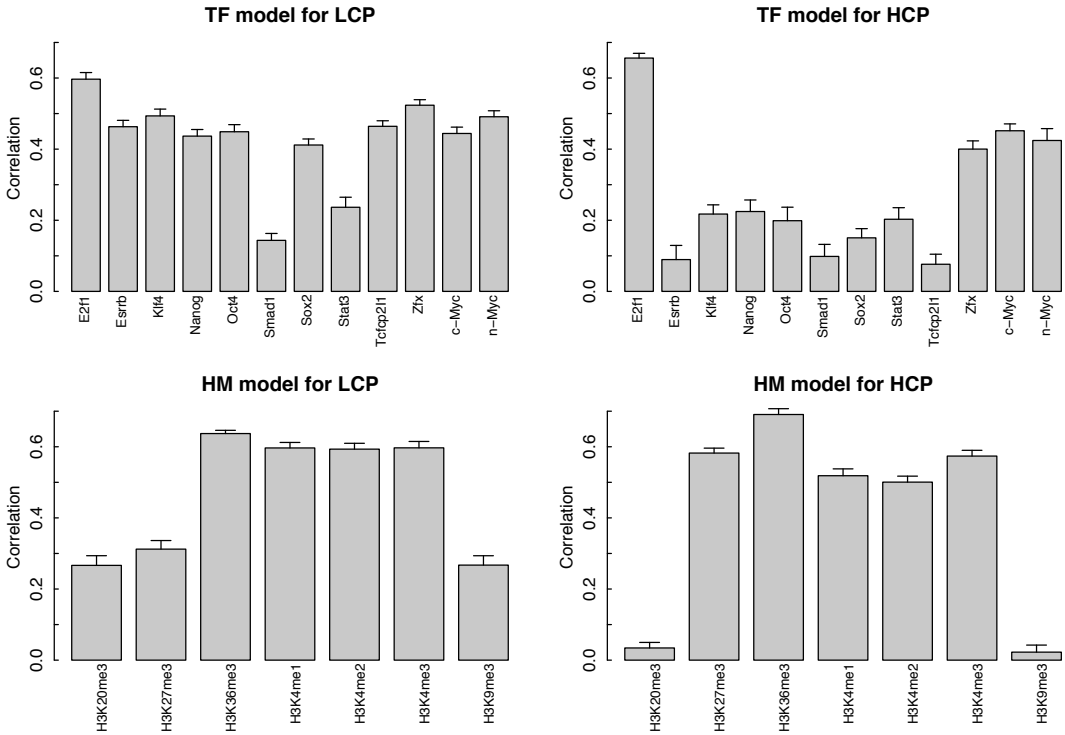

Figure S10: Prediction Accuracy of each individual TF or histone modification for predicting Low CpG content (LCP) and high CpG content (HCP) genes. For each individual TF or histone modification, the prediction accuracy was calculated based on the SVR model using its signal in 80 TSS bins (TF) or all of the 160 bins (HM).

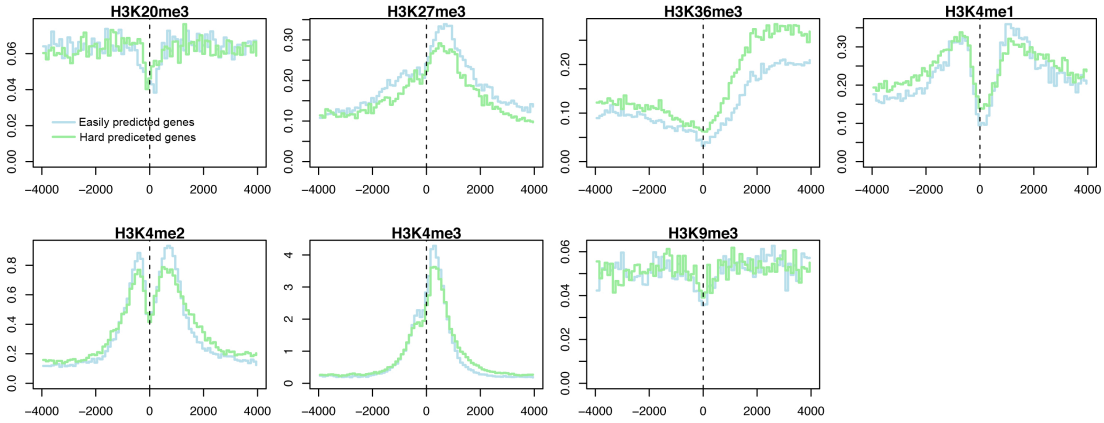

Figure S11: Histone modification patterns of 1000 most easily predicted genes versus 1000 hardest to predict genes. For each histone modification, the accumulative signals of the two gene groups around the TSS regions are shown.

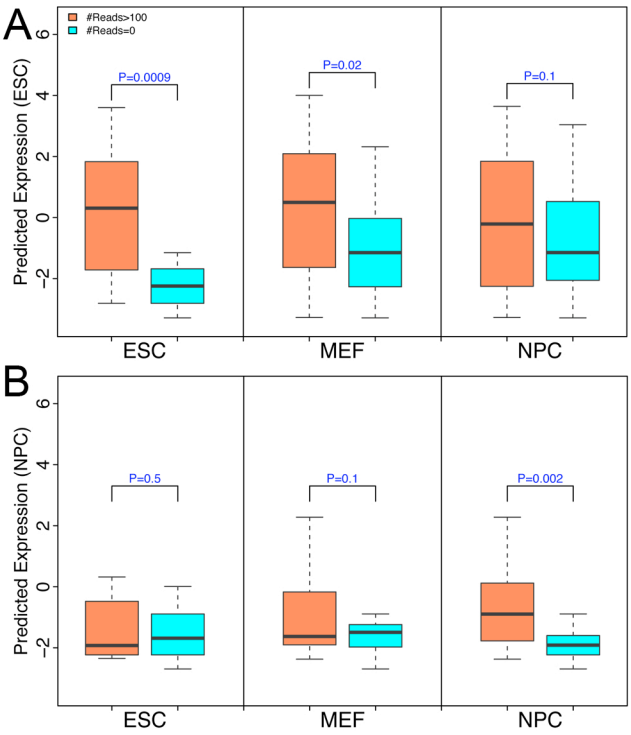

Figure S12: The HM model is predictive of microRNAs expression with cell line specificity. The same analysis in Figure 8 is performed, but only miRNAs that are not overlapping with any known genes are included. (A) Distribution of predicted microRNAs expression levels for highly and lowly expressed microRNAs in ESC (left), MEF (middle) and NPC (right). The model is trained on data for protein-coding genes in ESC. (B) Similar to (A), but the model is trained on data in NPC.

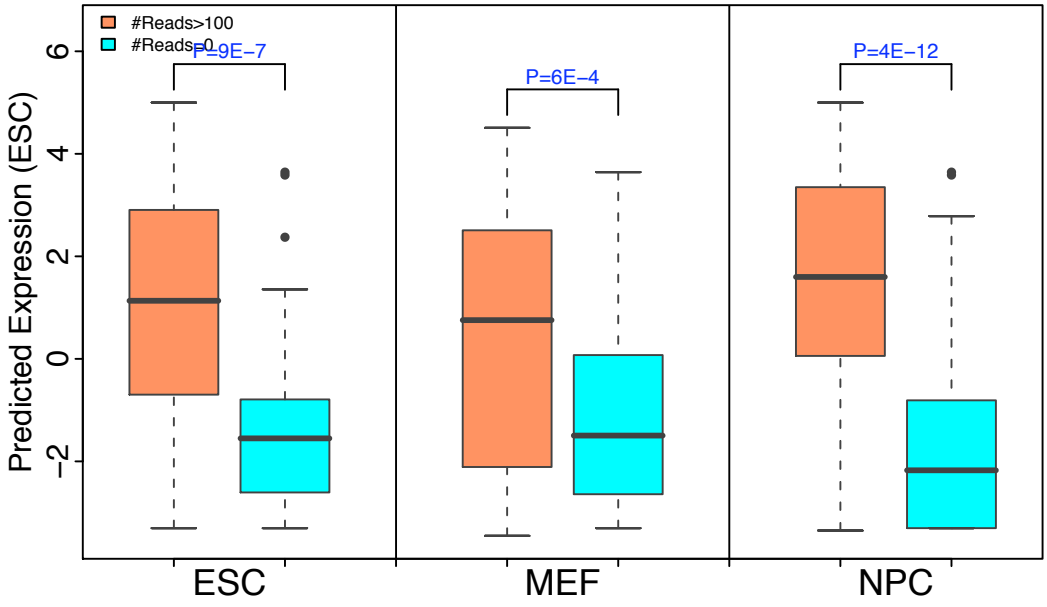

Figure S13: Transcription factor binding signal in microRNA promoters is predictive of microRNA expression levels with cell line specificity. The distributions of predicted microRNA expression levels for highly (with >100 mapped reads) and lowly (no mapped read) expressed microRNAs in ESC (left), MEF (middle) and NPC (right) are shown. The model is trained on data for protein-coding genes in ESC.

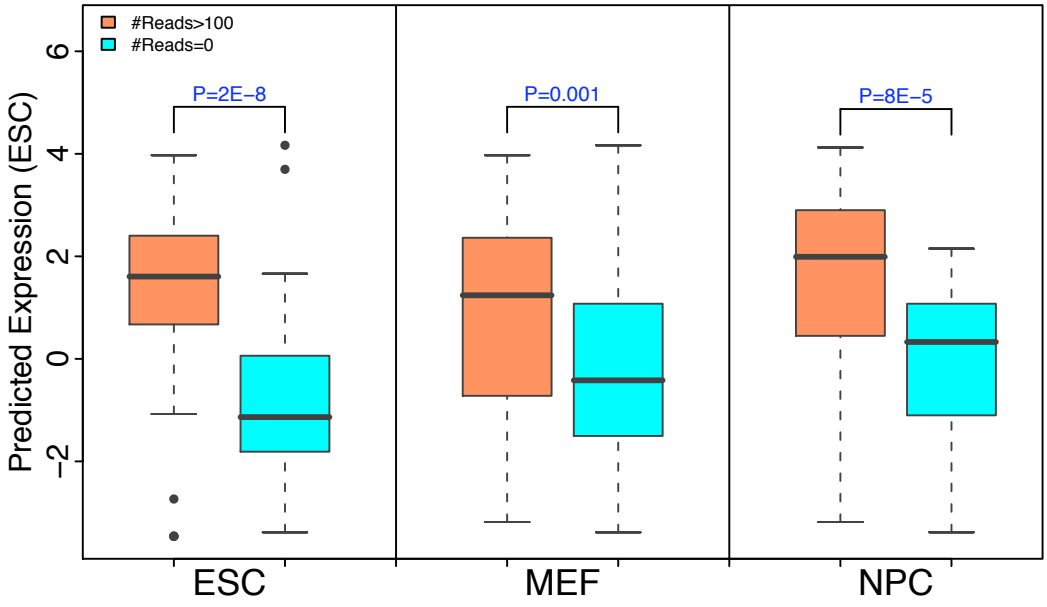

Figure S14: Histone modification pattern in microRNA promoters is predictive of microRNA expression levels with cell line specificity. The distributions of predicted microRNA expression levels for highly (with >100 mapped reads) and lowly (no mapped read) expressed microRNAs in ESC (left), MEF (middle) and NPC (right) are shown. The model is trained on data for protein-coding genes in ESC. Promoters of microRNA genes are based on computation analysis in [1].

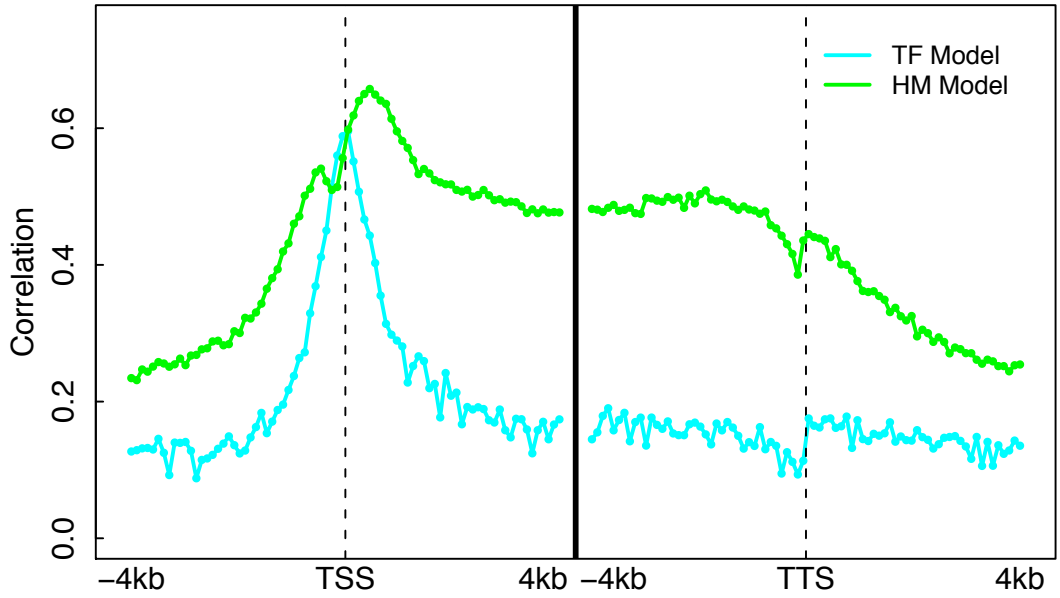

Figure S15: The accuracy of linear regression based TF and HM models for predicting gene expression in each of the 160 bins.

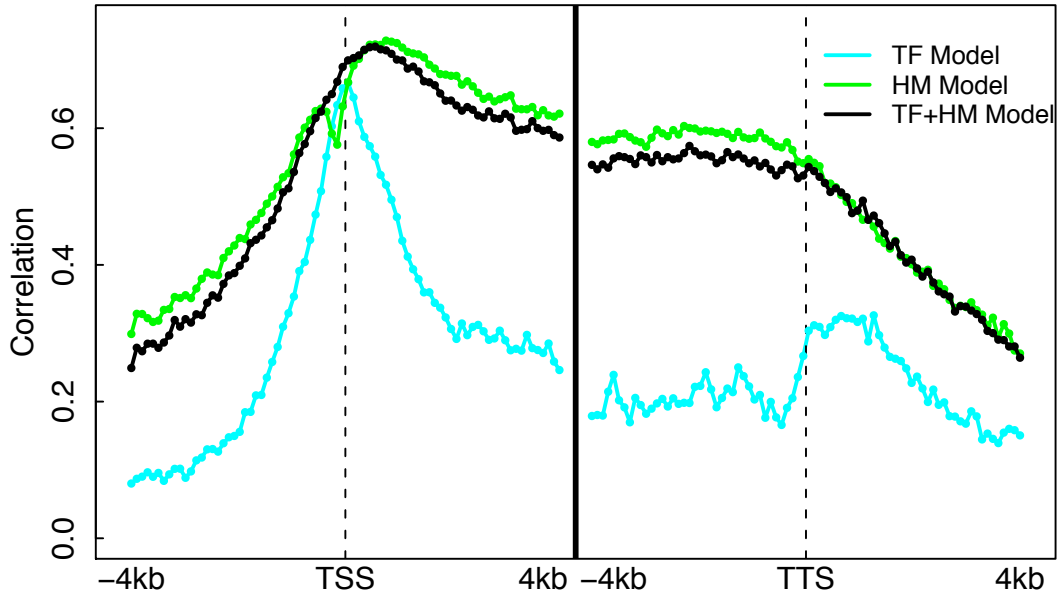

Figure S16: Prediction of accuracy of three models when applied to the ENCODE K562 data. The prediction accuracy of the TF model, the HM model and a combined TF+HM model in each of the 160 bins was shown.

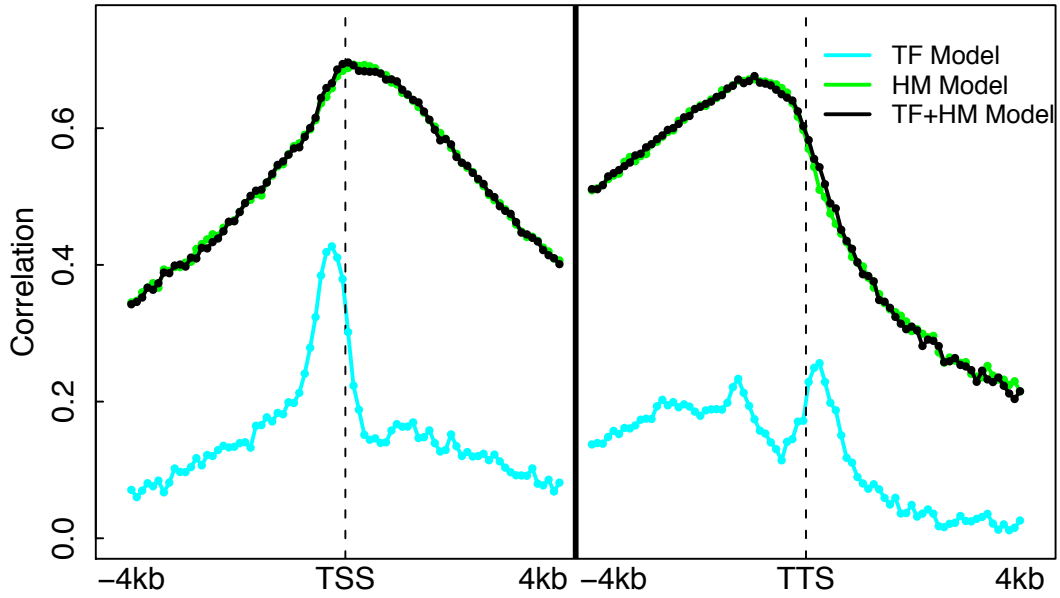

Figure S17: Prediction accuracy of three models when applied to the modENCODE early embryo data. The prediction accuracy of the TF model, the HM model and a combined TF+HM model in each of the 160 bins was shown.

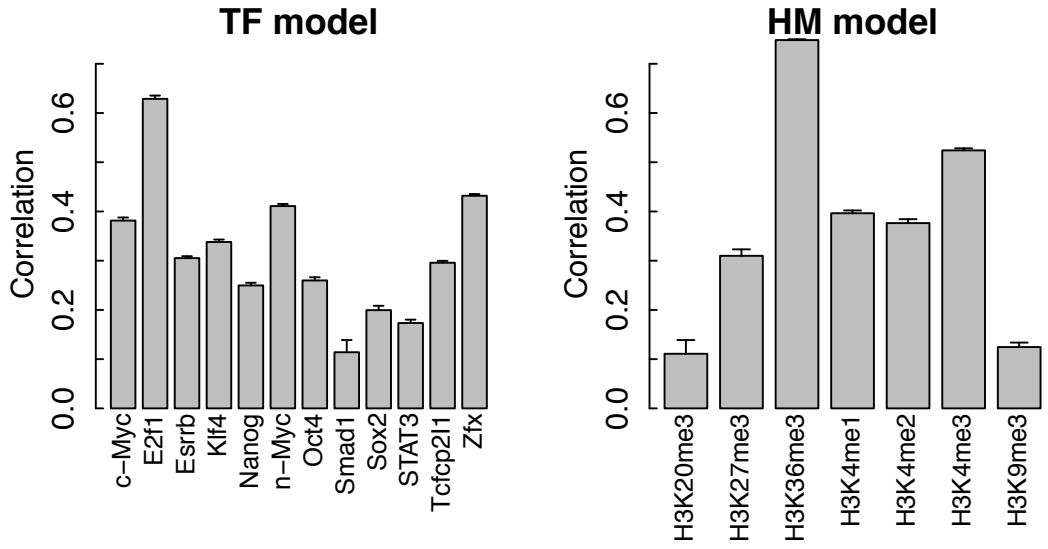

Figure S18: Prediction accuracy of individual TF binding profile (A) or histone modification (B) using average signals within the exonic regions.

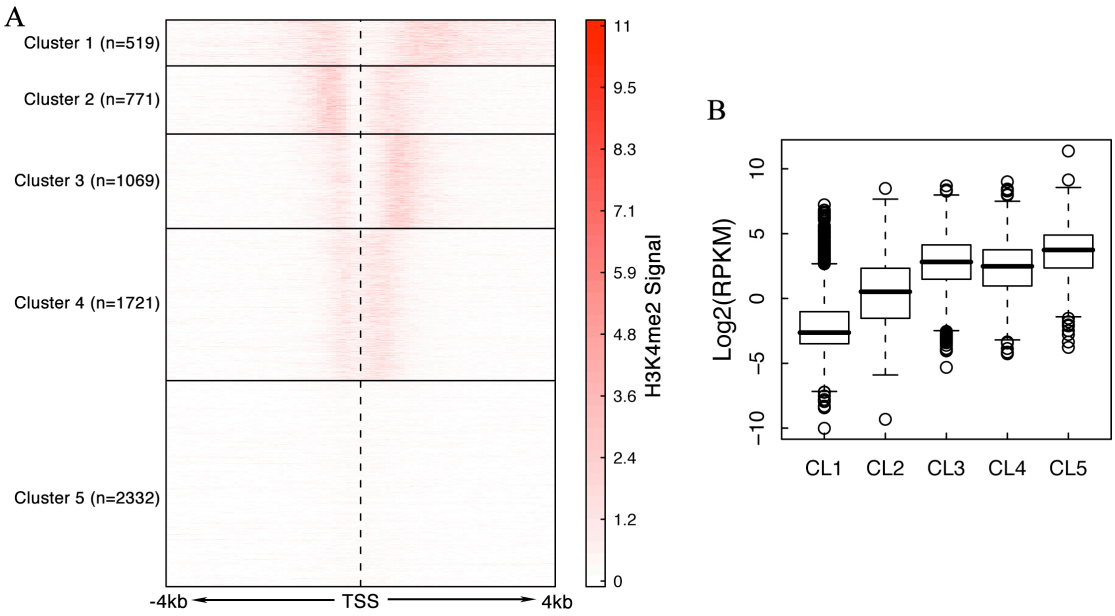

Figure S19: Clustering mouse genes into five clusters based on H3K4me2 profiles. (A) K-means clustering of H3K4me2 profiles in regions from −4 to +4 kb around the TSS of a set of nonoverlapping genes. The Euclidean distance was used as a measure of similarity. The number of genes within each cluster is shown in the parenthesis. (B) Expression levels of genes in each of the five cluster in the mESC cells.

Table S1: The number of peaks and target genes for each of the 12 TFs in mESC. Target genes were identified as those that have at least one binding peak centered within [-1kb, 1kb] around its TSS.

TF	#peak	#gene(-1500~500bp)
Smad1	1126	28
Sox2	4526	336
Stat3	2546	529
Nanog	10343	650
Oct4	3761	1017
c-Myc	3422	3011
Esrrb	21647	3140
Tcfcp2l1	26910	4483
Klf4	10875	4822
n-Myc	7182	5606
Zfx	10338	6511
E2f1	20699	10932

Table S2: Comparison of prediction accuracies by different models/methods. The accuracy is represented as the Pearson correlation coefficient between the predicted and the experimental values in the cross-validations.

Model	Two-layer SVR	SVR (best bin)	SVR (max signal)	SVR (avg signal)	MLR (best bin)	MLR (avg signal)[a]	PC regression[b]
TF	0.77	0.7	0.74	0.76	0.6	0.65	0.81
HM	0.82	0.7	0.82	0.81	0.66	0.76	NA
TF+HM	0.82	0.71	0.82	0.81	0.67	0.76	NA

a: multiple linear regression used in Karlic et al. 2010
b: method used in Ouyang et al. 2009