A statistical framework for modeling gene expression using chromatin features with application to modENCODE datasets

Chao Cheng[1], Koon-Kiu Yan[1], Kevin Y. Yip[1,4], Joel Rozowsky[1], Roger Alexander[1], Chong Shou[1], Mark Gerstein[1,2,3,$]

1 Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut, USA

2 Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, USA

3 Department of Computer Science, Yale University, New Haven, Connecticut, USA

4 Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong.

$ To whom correspondence may be addressed: pi@gersteinlab.org

Abstract

We develop a statistical framework to study the relationship between chromatin features and gene expression. This can be used to predict gene expression of protein coding genes, as well as microRNAs. We demonstrate the prediction in variety of contexts, focusing particularly on the modENCODE worm datasets. Moreover, our framework reveals the positional contribution around genes (upstream or downstream) of distinct chromatin features to the overall prediction of expression levels.

Introduction

In eukaryotes, nuclear chromosomes are organized into chains of nucleosomes, which are in turn composed of octamers of four types of histones wrapped around 147 base pairs of DNA. Modifications of these core histones are central to many biological processes including transcriptional regulation [1], replication [2], alternative splicing [3], DNA repair [4], apoptosis [5, 6], gene silencing [7], X-chromosome inactivation [8] and carcinogenesis [9, 10]. Among them, transcriptional regulation is one of the most important and thereby intensively investigated processes [1, 11, 12]. Histone modifications have been demonstrated to regulate gene transcription in positive or negative manners depending on the modification site and type [13-18]. For example, a genome-wide map of 18-histone acetylation and 19-histone methylation sites in human T-cells indicates that H3K9me2, H3K9me3, H3K27me2, H3K27me3 and H4K20me3 are negatively correlated with gene expression, whereas most other modifications including all the acetylations are correlated with gene activation [18, 19]. As an extreme case, histone modifications play critical roles in X-chromosome inactivation in females to equalize the expression of X-linked genes to those in male animals [19, 20]. Histone modifications are thought to affect transcription through two mechanisms: (1) modifying the accessibility of DNA to transcription factors by altering the local chromatin structure; and (2) providing

specific binding surfaces for the recruitment of transcriptional activators and repressors [11, 17, 21-23].

The large number of possible histone modifications has led to the "histone code" hypothesis, which states that combinations of different histone modifications specify distinct chromatin states and bring about distinct downstream effects [24-26]. Moreover, one histone modification may influence another by recruiting or activating chromatin-modifying complexes [27]. However, a study in yeast revealed only simple and cumulative functional consequences for combinations of histone H4 acetylation rather than a complicated synergistic histone code [28]. Two other studies, one in yeast and the other in *Drosophila,* also demonstrated that histone modifications are highly correlated with each other and are partially redundant in function [13, 17], presumably conferring robustness in relation to epigenetic regulation [29]. Alternatively, the high correlation between histone modifications may have been overestimated as a result of differences in nucleosome density or other unknown biases [29]. So far, knowledge about the effect of histone modifications on transcriptional regulation is still limited, and the degree of complexity of the histone code is far from clear. To further understand the relationship between histone modifications and gene expression, we require a systematic analysis that integrates histone modification maps with other genome-wide datasets.

The model organism encyclopedia of DNA elements (modENCODE) project was launched in 2007 for the purpose of generating a comprehensive annotation of functional elements in the *C. elegans* and *D. melanogaster* genomes [30]. By using recently developed genome-wide experimental techniques such as ChIP-chip, ChIP-seq and RNA-seq [31, 32], modENCODE has generated a large amount of data including gene expression profiles, histone modification profiles, and DNA binding data for transcription factors and histone-modifying proteins. This large compendium of datasets provides an unprecedented opportunity to investigate the relationship between chromatin modifications and transcriptional regulation using an integrative approach.

In this study, we endeavor to construct a general framework for relating chromatin features with gene expression. We apply a multitude of supervised and unsupervised statistical methods to investigate different aspects of gene regulation by chromatin features. Leveraging the rich data generated by the modENCODE project, we use *C. elegans* as a primary model to illustrate our formalism. Nevertheless, we tested the generality of our methods using a variety of species ranging to yeast to human. More specifically, we show that chromatin features can accurately predict the expression levels of genes and collectively account for at least 50% of the variation in gene expression. We also study the importance of individual features, examine the combinatorial effects of chromatin features, and investigate to what extent the histone code hypothesis is valid. By applying the chromatin-based model to predict the expression of coding genes and microRNAs at different developmental stages, we further address the developmental stage specificity of chromatin modifications and suggest that chromatin features regulate transcription of coding genes and microRNAs in a similar fashion.

As more and more genome-wide ChIP-Seq and RNA-Seq data are going to be generated via the modENCODE project and the ENCODE project [33] in the near future, the methods of data integration proposed in this work have various potential applications.

Results

Chromatin features show distinct signal patterns around genic regions

To systematically study the genome-wide properties of various chromatin features, we collected more than 50 ChIP-chip and ChIP-seq profiles of histone modifications and DNA binding factors in *C. elegans* from the modENCODE project (see methods). We divided the DNA regions around (+/- 4kb) the transcription start site (TSS) and transcription termination site (TTS) of each transcript into small 100-bp bins and calculated the average signal of the chromatin features in each bin. As a result, each bin was assigned a matrix whose elements are the average signals of different features in different transcripts (Figure 1). Figure 2A shows the rich spatial pattern of 16 features in early embryonic stage (EEMB), where the signals are averaged over all transcripts. We first observed that the upstream and downstream regions of TSS and TTS are clearly distinct. Most chromatin features have higher signals in the transcribed regions (downstream of TSS and upstream of TTS). Interestingly, we found that Pol II has the strongest binding signal in regions right after the TTS, rather than within the transcribed region (Figure 2A). The enriched binding signals right after the TTS may indicate the importance of anti-sense transcription as a regulatory mechanism for gene expression [14, 33]. Strong Pol II signal was also observed at regions before the TSS in some other developmental stages (Additional file 1, Supplemental Figure S1), which was also reported previously in *C. elegans* by [34], and in mouse and human is thought to be related to the accumulation of TSS-associated RNAs [35, 36]. The signal pattern of histone H3 suggests that nucleosomes have lower occupation density in regions around the TSS and TTS than within the transcribed regions. H3K4me2 and H3K4me3 are enriched upstream of the TSS, consistent with their reported role as histone marks for active promoters [14]. On the other hand, signals for H3K9me2 and H3K9me3 are depleted around TSS compared to neighboring regions, which may reflect the low density of nucleosomes around the TSS of genes [28].

Chromatin features exhibit distinct spatial correlation patterns with gene expression levels

The different chromatin features display distinct spatial patterns. It is thus worthwhile to explore the relationship between these patterns and the level of gene expression. Making use of RNA-seq data obtained from the different stages of C. *elegans,* we quantified the expression level of each gene. For each bin, we then calculated the correlation between the gene expression levels and the average signals of each chromatin feature of the bin. Figure 2B shows the spatial variation of these correlation coefficients around TSS and TTS. According to the correlation patterns, there are two main types of chromatin features: ones that

are positively correlated with gene expression (such as H3K79me1, H3K79me2 and H3K79me3) and ones that are negatively correlated with gene expression (such as H3K9me2 and H3K9me3). While some features show largely uniform correlations across the 16kb regions, some others are more variable across the regions. For example, H3K79me2 has a high correlation coefficient (0.65) near the TSS, but rather a low correlation (0.10) downstream of the TTS. It is interesting to observe that the negative features tend to have more uniform spatial patterns while the positive features tend to show greater variation. In addition, for chromatin features such as H3K79me2, although the average signal intensity decreases with distance downstream from TSS, the correlation between the feature signal and the expression level remains high. This pattern suggests that, while some chromatin features have the strongest average signals only at some highly specific regions, the differences of their signals between genes with low and high expression levels remain strong over much broader regions.

We chose the long window size of 4kb in order to inspect how fast the signals of the chromatin features fade out as we move away from TSS and TTS. Indeed, the correlations of some chromatin features (e.g., H3K9me3) remain strong a few kb away from the TSS and TTS, and the fading could only be observed at the 4kb boundaries. To make sure that our conclusions are not affected by short genes with some bins having both the identities of being within 4kb downstream of TSS and within 4kb upstream of TTS, we also did the correlation analysis only on transcripts longer than 8kb, and found that the correlation patterns are the same (Additional file 2, Supplemental Figure S2). Also, as the *C. elegans* genome is quite compact, the region 4kb upstream of a TSS or downstream of a TTS could be overlapping with another gene. We thus repeated the analysis using transcripts that are at least 4kb away from any other known transcripts, and again obtained similar correlation patterns (Additional file 3, Supplemental Figure S3). Furthermore, analysis based on bins within intergenic regions again resulted in a similar correlation pattern. Therefore, the high correlation of gene expression with feature signal at distant locations does reflect the long-range effects of their regulation, instead of an artifact caused by chromatin structure of the nearby genes.

Furthermore, to assess whether the trends we observed are universal to all developmental stages rather than specific to EEMB, we repeated the analysis in other stages including late embryo, larval stages and young adult. Although the exact values of correlation coefficients vary across stages, the spatial patterns are consistent in all stages (see Additional file 4, Supplemental Figure S4 for the results in L3). In addition, a large number of genes are associated with multiple transcripts corresponding to different alternative splicing isoforms. In many cases, the overlap between these transcripts is substantial, which might affect the correlation patterns between chromatin features and expression. We thus repeated the correlation analysis using only genes with a single transcript, and obtained the same qualitative results (Additional file 5, Supplemental Figure S5).

Among the chromatin features shown in Figure 2, MES-4 and MRG-1 are factors associated with X-chromosome inactivation [37, 38]. These factors are supposed to have different binding patterns in the X chromosome than in autosomes. We

therefore analyzed their correlation patterns in X genes and autosomal genes separately. As expected, we found that MES-4 and MRG-4 associate predominantly with autosomal DNAs, while the DCC subunits bind specifically with X-chromosomal DNAs (data not shown), which is in line with previous reports [19]. Consistent with this finding, MES-4 and MRG-4 show stronger positive correlation with autosomal gene expression.

Unsupervised clustering reveals general activating and repressing chromatin features for individual genes

As some chromatin features are positively correlated with gene expression levels and some are negatively correlated, the two groups potentially represent general active and repressive marks of gene expression. Yet since these correlations capture only the average behavior across all genes, it is still not clear if these features are strong indicators of the expression levels of individual genes.

In order to examine the relationship between chromatin features and the expression levels of all individual genes, we performed a two-way hierarchical clustering of both the chromatin features and the annotated genes, according to the feature signals at the TSS bins (Bin 1). As shown in Figure 3A, genes can be divided into two clusters (labeled as H and L respectively) based on the signals of the 16 features. We found that the two clusters roughly correspond to genes with high expression levels (H) and genes with low expression levels (L), respectively (Figure 3B). These two clusters are characterized by complementary patterns of chromatin features. Cluster H is characterized by high signals of 11 features (the right component of the upper dendrogram), and low signals for the other 5 features. We note in particular that highly expressed genes tend to have a strong H3K36me3 signal, which is consistent with the role of H3K36me3 as a chromatin mark that activates transcription of associated genes. Similarly, the well-known repressive mark H3K9me3 shows a low signal. Compared to cluster H, genes in cluster L show the opposite pattern of chromatin signals.

To explore which regions around the TSS and TTS provide the greatest power in determining gene expression levels, we repeated the two-way clustering procedure for each of the 160 bins around TSS and TTS. Figure 3C shows the resulting t-statistics. We observe that the signals slightly downstream of TSS are the most informative. In general, the t-statistics decrease as the distance from TSS or TTS increases. The decay is steeper at the region downstream of TTS.

The above integrative analysis involves all chromatin features. To examine how each feature individually affects gene expression, for each feature we performed hierarchical clustering of the genes based on the collective signals of the feature at all 160 bins. An example is shown in Figure 3D, in which signals of the single feature H3K79me2 at the different bins were used to cluster the genes. As in the case when all chromatin features were used, the signals from single chromatin features can divide genes into two clusters (that are not exactly the same as, but similar to, the ones obtained from all features) with a significant difference in expression level (Figure 3E). Again we quantified the power of each feature in distinguishing genes with high and low expression levels using t-statistics. As

shown in Figure 3F, apart from a few exceptions (black bars), most features are informative. The most informative features are H3K79me2, H3K79me3 and H3K4me2. The informative features can be further grouped into two classes. Activating features are those that are positively correlated with gene expression (cyan) and repressive features are those that are negatively correlated (blue).

Chromatin features can statistically predict gene expression levels with high accuracy using supervised integrative models

The above analyses suggest that gene expression levels can be at least partially deduced from chromatin features. To examine how much of gene expression is determined by chromatin features, we tried to predict gene expression levels using the features. We started with the simplified task of distinguishing highly-expressed and lowly-expressed transcripts, where the two classes of transcripts were constructed by discretizing gene expression levels (see Methods). We divided all the transcripts into training and testing sets, and learned a support vector machine (SVM) model from the signals of all 13 chromatin features of the training transcripts at a certain bin (Figure 1). The model was then used to predict to which class each transcript in the testing set belongs. We repeated the procedure for all 160 bins, and 100 different random splitting of the transcripts into training and testing sets for each bin (see Materials and Methods). We represented the overall performance of the model using the Receiver Operating Characteristic (ROC) curve and further quantified the accuracy using the area under the curve (AUC). Figure 4A shows the ROCs corresponding to the prediction performance of five different bins. As compared to random ordering, which would give a diagonal ROC curve on average with an expected AUC of 0.5, we observed that all five curves are much better than random but with diverse performance, which indicates that all the bins are useful to classify gene expression but they are not equally informative. This result is consistent with what we have observed using the unsupervised method described above (Figure 3F). Instead of using SVM, we also learned support vector regression (SVR) models using similar procedures (see Methods) to predict expression values directly. Figure 4B shows that there is a high positive correlation (0.75) between the predicted levels from an SVR model and the actual expression levels measured by RNA-seq. This analysis suggests that chromatin features explain at least 50% of gene expression variation (see Methods).

We then compared the prediction accuracy of all 160 SVM models learned from the different bins. As shown in Figure 4C, the models learned from regions around the TSS (-300 to 500 bp) and upstream of the TTS (-200 bp to 0 bp) have highest accuracy with AUC values greater than 0.9. Prediction accuracy decreases gradually as we move away from these regions, which confirms the spatial effects that we observed from the unsupervised analysis (Figure 3C).

We have also tested more comprehensive models that combine the chromatin features in 40 bins around the TSS (-2kb~2kb). These comprehensive models achieve slightly higher prediction accuracy than those based on single bins, yet the enhancement is not dramatic with average AUC of 0.94 for the classification

model (SVM) and average correlation coefficient of 0.75 for the regression model (SVR)(Additional file 6, Supplemental Figure S6).

We then learned SVM models using only features of individual types. As shown in Figure 5A, the AUC obtained by using all features (black) is comparable to the AUCs obtained from models using only particular subsets of features. Strikingly, the model involving only the 9-histone modification features is almost as accurate as the model involving all 16 features. We further divided the histone modification features into 4 subsets: modifications on K4, K9, K36 and K79, respectively. While the integrated model with all histone modifications achieves an AUC value of 0.9, using just one of the subsets can yield an AUC higher than 0.8 (see Figure 5B). In particular, the set H3K79 is found to be most predictive, which again confirms our previous finding of the importance of these histone modifications in regulating gene expression (Figure 3F).

The results of the supervised analysis suggest that chromatin features are not only correlated with expression, but are also predictive of the expression levels of individual genes with good accuracy and could explain a large portion of the expression differences between different genes. We note that histone modifications may have other regions of enrichment that are informative about gene expression: for instance, the percentage of gene length with strong histone modification signals. We therefore examined the power of using these features for predicting gene expression levels. Specifically, we calculated the percentage of transcribed regions with strong signals (>10%) for all genes. Using them as predictors, we obtained high prediction accuracy (AUC=0.90). However, a combination of these percentage features with the original chromatin features does not lead to obvious improvement in prediction accuracy, indicating that they are redundant.

Combination of chromatin features contribute to gene expression prediction

Both the unsupervised and supervised analyses above suggest that chromatin features possess a certain level of redundancy. In the unsupervised clustering (Figure 3A), different chromatin features show similar signal patterns around the TSS regions of genes. In the supervised predictions (Figure 5), high accuracy was achieved by multiple features as well as feature subsets. Though the SVR model offers good prediction power, it may be instructive to build a simpler linear regression model to explore to what extent the chromatin features are redundant, and to what extent they are interacting in a combinatorial fashion. Specifically, for each bin, we modeled the expression level y as a linear combination of the effects of individual histone modification features x_i and their products $x_i x_j$: $y \sim \sum x_i + \sum_{i<j} x_i x_j$. We found that among the 66 (12×11/2) possible interactions between the 12 distinct histone modification features, many interactions are statistically significant. For example, for bin 1, we detected 12 significant interactions (P<0.001, linear regression) between the histone modifications (Additional file 7, Supplemental Table S7).

To quantify the importance of these interactions in determining gene expression levels, we compared the above regression model with a singleton model that does not contain the interaction terms: $y \sim \sum x_i$. By evaluating the prediction power of the two models using a cross-validation method, we found that with respect to the singleton model the interaction model improves prediction accuracy by 4%. Thus, the contribution of interactions among chromatin features to gene expression prediction is not substantial.

We further examined each pair of modifications individually to see if any of modifications is redundant of another one. Using simplified models each involving only two modification features, we found that no two histone modifications are completely redundant (Additional file 8, Supplemental Table S8). These results were confirmed by a similar analysis based on mutual information (Additional file 9, Supplemental Figure S9). Two examples are shown in Figure 6. In each example, we considered a specific pair of histone modification features, and divided all genes into four categories based on the signals of the two features at their TSS bins. In the first example (Figure 6A), expression levels are the lowest when both H3K4me3 and H3K36me3 are low but moderate if either one of them is high. This suggests that both features are activators. When both features have high signals, an even higher expression level is observed, showing that the two are not totally redundant. In the second example (Figure 6B), H3K9me3 is found to repress gene expression in general, while H3K79me3 is found to activate gene expression. As expected, a combination of high H3K9me3 signal and low H3K79me3 signal results in a lower expression level than when both signals are low. When the signals of both features are high, we observe a significant difference in gene expression compared to the other 3 cases, indicating that the features contribute to gene expression regulation in a collective manner.

Our analyses of the interactions between the above chromatin features only considered binary interactions between two features. For higher-order relationships involving more features, it is infeasible to perform the same type of analyses, as the number of feature combinations would become intractable. Also, the above analyses only suggest which features interact with each other, but do not explain how the features interact. In particular, the complex correlations between features and gene expression make it difficult to extract directional relationships between them (Additional file 10, Supplemental Figure S10A). We therefore used Bayesian networks to study the higher order relationships between the chromatin features and gene expression (see Additional file 11 for details).

The chromatin model is developmental stage-specific

We have previously constructed an integrative model using chromatin features at the EEMB stage of *C. elegans* development and used it to predict gene expression levels at the same stage. How well can we predict gene expression levels at other developmental stages using the chromatin feature data from

EEMB? To answer this question, we applied the model to predict gene expression at EEMB, L1 (larva stage 1), L2, L3, L4, and adult. Specifically, the chromatin feature data from EEMB were combined with expression data from a stage to train a SVM model, which was then used to predict gene expression levels of other genes at that stage. As shown in Figure 7, the chromatin model based on EEMB data is able to predict the expression at other developmental stages with reasonable accuracy (AUC=0.8). However, the predictions of gene expression levels in all these stages have lower accuracy than the predictions for EEMB itself. This result suggests that signals from chromatin features are developmental stage-specific and regulate biological processes in a dynamic manner depending on the particular stage. The stage specificity is more apparent, when we apply the model to genes that are differentially expressed between stages. For example, we have identified 4,042 genes that differ in expression levels by at least 4-fold between EEMB and L3. Using the EEMB chromatin model to predict the expression level of these genes, the prediction accuracy further decreases (AUC=0.70).

Chromatin features show different correlation patterns with different genes in an operon

In *C. elegans* some neighboring genes are organized into operons. The genes in an operon are co-transcribed as a polycistronic pre-messenger RNA and processed into monocistronic mRNAs [39, 40]. Here we investigate the differential signals of chromatin features among genes in operons and how this organization affects their expression levels. We collected the first, second and last genes in 881 *C. elegans* operons and calculated the signals of chromatin features in each of the 160 bins around their annotated TSS and TTS. We observed strong correlations between expression levels and chromatin feature signals for the first genes (Figure 8). In comparison, the correlation patterns for the second and last genes of the operons are not as apparent (see Additional file 12, Supplemental Figure S12 for all features). The weaker correlations could be caused by the lack of signals for some histone modification types. As we observed, the mark for active promoters, H3K4me3, demonstrates strong signals around the TSS of the first genes, which is the shared promoter of genes in the same operon. In the upstream region of the internal genes, the H3K4me3 signal is often relatively weak. Alternatively, the weak correlation for internal genes may also be explained by the intensive post-transcriptional regulation of these genes, which can not be captured by our chromatin feature based model [41]. In fact there is only weak correlation (PCC=0.10) between the expression levels of the first and the second genes. Moreover, on average the first genes are two-fold and three-fold more highly expressed than the second genes and the last genes, respectively. Taken together, although genes in the operons are co-transcribed, they are regulated post-transcriptionally to achieve distinct expression levels [41].

Chromatin models learned from protein-coding genes are able to predict microRNA expression levels with high accuracy

Do chromatin features influence transcription of microRNAs in the same way as they do with protein-coding genes? As a way to study the similarity of the two mechanisms, we investigated the effectiveness of the chromatin model learned from protein-coding genes in predicting microRNA expression. Since precise transcription start sites are not available for most worm microRNAs, we calculated the signals of chromatin features in the genomic regions corresponding to pre-microRNAs, and used them as the input features for our chromatin model.

We predicted the expression levels of 162 worm microRNAs with genomic locations obtained from miRBASE [42]. We then compared our predictions with the experimental measurements performed by Kato et al [43]. As shown in Figure 9, our predictions are in good agreement with the experimental results in the EEMB stage (see also the prediction results for the L3 stage in Additional file 13, Supplemental Figure S13). Some microRNAs locate within or nearby gene loci, which may confound the prediction of microRNA expression. To address this issue, we also checked the prediction accuracy using only microRNAs that are away from any known gene, and obtained similar prediction accuracy (PCC=0.62).

It is interesting to see that the expression of microRNAs can be accurately predicted using a chromatin model trained by data for protein-coding genes. Consistent with previous report on microRNA transcriptional regulation [44, 45], this result suggests that microRNAs and protein-coding genes share a similar mechanism of transcriptional regulation by chromatin modifications.

As with the prediction of expression levels of protein-coding genes, the prediction accuracy of microRNA expression also shows developmental stage specificity. When the signals of the chromatin features from EEMB stage were used, the resulting model achieved the best accuracy when predicting microRNA expression at the same stage (PCC=0.60), whereas for stages L1, L2, L3, L4 and adult, the accuracy is much lower (PCC<0.50) (Additional file 14, Supplemental Figure S14). Similarly, when chromatin features at L3 were used to train the model, the model achieved better prediction results in L3 than in other stages.

Application to other organisms

The models described above provide a useful tool to integrate gene expression and chromatin data. Currently the *C. elegans* dataset is the best one to demonstrate the utility of the method and we have focused on it here. However, we know that further integrated genomic datasets (comprising matched genome-wide histone features and expression measurements) are coming in many other organisms. Thus, to illustrate the broad utility of our method, we demonstrate here how readily it can be applied in other contexts. Specifically, we have packaged our methods as a tool and applied it to data sets from four other organisms: yeast, fruit fly, mouse and human. The results indicate that chromatin features, in particular histone modifications, are highly correlated to gene expression levels in all these organisms (Figure 10). More importantly, the relative statistical contribution of each histone modification type to expression is

similar in tested organisms (and also in different tissues, cell-lines, and developmental-stages). For example, H3K4me3 signals around TSS of genes show high predictive capability in all the analyses we have performed. We also found that the models based on expression levels measured by RNA-seq achieved higher prediction accuracy than those by microarrays, consistent with the higher measurement accuracy of RNA-seq compared to microarrays. Our method can, of course, be applied to multiple data sets in each species (e.g. different developmental stages in fruit fly). Figure 10 shows only a single illustrative example for each species. We only show initial statistical analysis here, further biological interpretation would, of course, be the subject of future studies.

Discussion

In this study, we have presented a systematic analysis on the genome-wide relationship between chromatin features and gene expression. We have shown that, in terms of gene expression prediction, information from different histone modification features is considerably redundant. Here in this paper, we use the modENCODE worm data to exemplify our analysis. In fact, we have applied our methods to two other histone modification data sets: the human CD4+ T-cell data [46] and mouse embryonic stem cell deta [47]. In both data sets, we found that histone modifications account for more than 50% of variation of gene expression and distinct modification types were redundant for predicting gene expression levels. This is consistent with a recent study by Karlic et al. performed in human CD4+ T-cells [48].

The existence of a "histone code" has been intensively debated since the time that the hypothesis was first proposed ten years ago [24, 25]. Previous studies have demonstrated both pros and cons for the hypothesis [11, 28, 49, 50]. Indeed, for some specific genes, it has been demonstrated that the patterns of a subset of histone marks could be viewed as an accurate predictor of gene regulation in non-trivial manners [50]. Nevertheless, the readout of these patterns is largely gene specific and dependent on the cellular context, which makes it difficult for these cooperative effects to be viewed as a universal "code". Therefore by using the term histone code, we might have underestimated the complexity and over-generalized the meaning of chromatin modifications and their roles in biological processes. On the other hand, at a global level, previous studies have reported substantial correlations among distinct chromatin features [13, 14, 17, 28, 51]. These results, and the information redundancy we observed, are consistent with the simple "histone code" argument [28], in which the combinatorial effects are cumulative rather than synergistic.

We have shown that chromatin features are strongly correlated with gene expression. Nevertheless, it should be noted that our models could not reveal if histone modifications are the "cause" or "consequence" of transcription. In fact, both directions of causality have been previously reported. Some studies have proposed that some histone modifications are the memory of past transcriptional events resulting from previous active transcription [52-54]. For

instance, it has been shown that phosphorylation in the tail of RNA pol II is required for H3K4me3, revealing that it is a direct consequence of Pol II passing through the TSS [55]. Other studies, however, have shown that chromatin modification changes precede changes in gene expression [56]. A recent study in human T-cells suggested that for both protein-coding and miRNA genes, activating histone marks were already in place before induction of expression, and these marks were maintained even after the genes were silenced [45]. This finding shows that histone modification can be both cause and consequence of gene transcription, and that a full explanation will require incorporation of additional data. Generalizing our model to follow a time course of changing histone modifications might be helpful for understanding this issue.

The supervised chromatin model trained from expression data for protein-coding genes can accurately predict the abundance of both protein-coding genes and microRNAs, which suggests that microRNAs and protein-coding genes share similar mechanisms of transcriptional regulation by chromatin modifications [44, 45]. To predict the expression levels of microRNAs, we used the signal of chromatin features around the start sites associated with pre-microRNAs, which might be several kilo-bases from the actual transcription start site of microRNA genes. Despite this caveat, our model still achieved high prediction power. We expect to obtain more accurate predictions if more precise annotation for microRNA genes becomes available in the future.

In summary, we have presented a series of supervised and unsupervised methods for analyzing multiple aspects of the regulation of gene expression by chromatin features. Apart from predicting gene expression, these methods can be used to address important biological questions such as combinatorial regulation and miRNA transcription. These and other statistical methods will be essential to gaining new understanding of biological processes from the tremendous amount of data that will soon be made available by large collaborative projects such as modENCODE.

Methods

Datasets and gene annotation

Expression levels for all annotated worm transcripts at different stages of development, including early embryo (EEMB), mid-L1, mid-L2, mid-L3, mid-L4 and young adult, were quantified using RNA-seq. RNA polymerase II binding across the genome at different stages was profiled using ChIP-seq. All the other chromatin features were profiled using ChIP-chip experiments. These chromatin features include histone H3 occupation, histone methylations (H3K4me2, H3k4me3, H3K9me2, H3k9me3, H3k27me3, H3K36me2, H3K36me3, H3K79me1, H3K79me2 and H3K79me3), binding of dosage compensation complex (DCC) proteins (SDC2, SDC3, DPY27, DPY28 and MIX1) and other X-chromosome inactivation factors (MES4 and MRG1). For some chromatin features such as H3K9me3, biological replicates using different antibodies were available. Profiles of these chromatin features were measured for different developmental stages, in particular at EEMB and L3. A list of the data, with their GEO IDs can be found in the supplemental materials (Additional file 15, Supplemental Table S15).

All these data are available from the modENCODE website at http://www.modencode.org. Operon information for *C. elegans* was obtained from a previous study by Blumenthal et al. [39]. The dataset contains a total of 881 operons with 2.6 genes in each of them on average.

MicroRNA expression levels at different developmental stages of *C.elegans* were obtained from small RNA-seq measurements performed by Kato et al. [43]. Annotation of worm transcripts was downloaded from WormBase at http://www.wormbase.org [57]. Annotation of nematode microRNAs was downloaded from the microRNA database miRBASE at http://www.mirbase.org [42]. Assembly version WS180 of *C.elegans* was used for gene and microRNA annotations and data processing of all the chromatin features.

Binning DNA regions

We obtained the genomic locations and structures of 27,310 protein-coding transcripts of *C.elegans* from WormBase. The contribution of each chromatin feature to gene expression is thought to be affected by many factors, in particular its position relative to the transcription start site (TSS). We therefore divided the DNA region from 4kb upstream to 4kb downstream of the TSS of each transcript into 80 small bins, each of 100 bp in size. The DNA region around the transcription termination site (TTS) of each transcript was also divided into 80 100-bp bins. For each bin, we calculated the average signal of each chromatin feature across all transcripts. Specifically, for chromatin features profiled by ChIP-chip experiments, the signals of the probes that fall into a bin were averaged. For features profiled by ChIP-seq experiments, the number of reads that cover a bin was counted and weighted according to their overlap with the bin. We note that for short transcripts less than 8kb in length, some bins around TSS and TTS overlap, and for transcripts representing alternative splicing isoforms of the same gene or located close to each other in the genome, their bins can also overlap. To ensure these issues do not affect our main findings, we have performed analysis using only genes that are longer than 8kb and genes that are far away from coding genes (see main text). It should also be noted that the precise TSS and TTS of worm transcripts are largely unknown and the locations used here usually represent the start and end positions of the protein-coding regions.

Hierarchical clustering

The data processing described above results in a matrix $A_{n\times m}$ for each of the 160 bins, where n is the number of transcripts and m is the number of chromatin features. To make the signals for different chromatin features comparable, we normalized the columns of A by subtracting the median and then divided by the standard deviation of each column across all transcripts. We performed hierarchical clustering analysis using the normalized matrix for a given bin. To evaluate the capability of a bin to discriminate between genes with high and low expression levels, we divided the transcripts into two clusters by splitting the resulting hierarchical tree at the top level. The expression levels of transcripts in the two clusters measured by RNA-seq experiments were compared using t-test. We repeated this procedure for all 160 bins, which resulted in a t-score for each

bin. Those t-scores reflect the capability of chromatin features in these bins to separate genes with low and high expression levels.

Similarly, given a specific feature we performed hierarchical clustering using its signals across all 160 bins. The clustering analysis was conducted for all chromatin features, and the capability of each feature to predict gene expression was evaluated and compared by their t-scores calculated as described above.

Supervised models for gene expression prediction

We constructed supervised learning models to integrate the chromatin features for gene expression prediction. In principle, the chromatin features of each of the 160 bins could contribute to regulation of gene expression. We therefore constructed the model in a bin-specific manner to investigate the relative importance of each bin for regulation of gene expression. We devised both classification and regression models, implemented by using the support vector machine (SVM) and support vector regression (SVR) [58] methods respectively.

In the classification model the expression levels of transcripts at a particular developmental stage (measured by RNA-seq and quantified as RPKM, reads per kilobases per million mapped reads) were discretized into two classes, with high and low expression level respectively, by setting the median expression levels as the cut-off values. The chromatin features in a given bin were then used as classifiers to predict the two classes. The prediction power of the classification model was evaluated using cross-validation. Specifically, we split the whole dataset into two halves, the training data and the testing data. The SVM model was first trained on the training data and then used to predict the classes of expression levels of the transcripts in the testing data. The predicted classes at various thresholds were compared with their actual classes to calculate the sensitivity (also called true positive rate, the proportion of actual positives which are correctly identified) and specificity (also called true negative rate, the proportion of negatives which are correctly identified). The tradeoff between sensitivity and specificity can be best visualized as a graphical plot of the sensitivity against 1-specificity, which is called a ROC (receiver operating characteristic) curve. The area under the ROC curve (AUC) is a frequently used summary statistic for measuring the prediction power of classification models.

In the regression model, we directly predicted the expression levels of transcripts rather than classifying them into two broad expression categories. The prediction power of the regression model was also checked using cross validation. The SVR model was trained on the training data and applied to the testing data. Then the predicted expression levels for transcripts in the testing data were compared with their actual levels measured by RNA-seq experiment. The correlation between predicted and actual expression level indicates the prediction power of the model.

In a linear regression model, the square of the correlation (R^2) between the predicted values and the actual values is equal to the fraction of total variance in the observed data explained by the predictions. We used this quantity to

estimate how much variation of gene expression can be explained by the chromatin features.

To estimate the predictive power of classification and regression models for each of the 160 bins, we repeated the cross validation procedure 100 times. The mean and standard deviation of the resulting 100 AUC scores were calculated for each bin as a measurement of the predictive power of the SVM classification model. Similarly, the accuracy of the SVR model for a bin was reflected by the mean and standard deviation of the 100 correlation coefficients.

Detecting combinatorial effects of chromatin features using linear models

To investigate the interaction between chromatin features, we constructed and compared the following two linear models: $y \sim \sum x_i + \sum_{i<j} x_i x_j$ (Interaction model) and $y \sim \sum x_i$ (Singleton model). The Interaction model takes into account the interaction terms. Based on the Interaction model, we identified significant interactions in each bin.

The power of the two models for predicting gene expression was evaluated by cross-validation. Data were randomly split into training and testing data sets. The models were trained on the training model and then applied to the testing data for validation. The accuracy of the models was measured by the correlation between predicted expression levels and experimental measurement.

To investigate the interactions among pairs of chromatin features, we constructed the simplified models involving only two features: $y \sim x_i + x_j + x_i x_j$. A significant interaction term would indicate that the interaction between the two features has a significant effect on gene expression.

Predicting expression levels of microRNAs

We downloaded the annotation of 162 *C.elegans* microRNAs from the miRBASE database [42]. For most microRNAs, the annotation provides no information about the transcription start sites. Instead, only the start and end positions of the corresponding pre-microRNAs (about 100 nt in length) are available. To predict the expression levels of microRNAs, we calculated the signals of all chromatin features within the associated pre-microRNAs and applied our model trained on chromatin features associated with protein-coding genes. We applied both the SVM classification and the SVR regression models to predict microRNA expression. The resulting predictions were validated using measured microRNA expression levels from small RNA sequencing performed by Kato et al [43].

Data sets for other organisms

In yeast, the expression levels of genes were measured by microarrays and available from Wang et al [59]; the histone modification data are performed by Pokholok et al [60]. In fruit fly, the gene expression and chromatin data at 12 different developmental stages was obtained by using RNA-seq and ChIP-seq

experiments, respectively, which are available from the modENCODE website at http://www.modencode.org. In mouse, the expression data in ESC (embryonic stem cell) and NPC (neural progenitor cell) cells were performed by Cloonan et al. [61]; and the histone modification data for matched cell lines were obtained from Mikkelsen et al [47] and Meissner et al [62]. In human, the gene expression data in K562 and GM12878 cell lines were performed by Mortazavi et al [63], and chromatin data were download from the ENCODE project at http://genome.ucsc.edu/ENCODE [2].

Availability of our code

All the analysis described in this paper was performed using the R package. The related R code and example data sets are available for download from http://archive.gersteinlab.org/proj/chromodel/index.html.

Acknowledgments

This work was supported by the NHGRI modENCODE project and the AL Williams Professorship funds. We thank Jason Lieb, Robert Waterston and Frank Slack for their comments and suggestions.

References:

1. Li B, Carey M, Workman JL: **The role of chromatin during transcription**. *Cell* 2007, **128**(4):707-719.
2. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE *et al*: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project**. *Nature* 2007, **447**(7146):799-816.
3. Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM, Misteli T: **Regulation of alternative splicing by histone modifications**. *Science*, **327**(5968):996-1000.
4. van Attikum H, Gasser SM: **The histone code at DNA breaks: a guide to repair?** *Nat Rev Mol Cell Biol* 2005, **6**(10):757-765.
5. Ahn SH, Cheung WL, Hsu JY, Diaz RL, Smith MM, Allis CD: **Sterile 20 kinase phosphorylates histone H2B at serine 10 during hydrogen peroxide-induced apoptosis in S. cerevisiae**. *Cell* 2005, **120**(1):25-36.
6. Cheung WL, Ajiro K, Samejima K, Kloc M, Cheung P, Mizzen CA, Beeser A, Etkin LD, Chernoff J, Earnshaw WC *et al*: **Apoptotic phosphorylation of histone H2B is mediated by mammalian sterile twenty kinase**. *Cell* 2003, **113**(4):507-517.
7. Schuettengruber B, Chourrout D, Vervoort M, Leblanc B, Cavalli G: **Genome regulation by polycomb and trithorax proteins**. *Cell* 2007, **128**(4):735-745.
8. Brinkman AB, Roelofsen T, Pennings SW, Martens JH, Jenuwein T, Stunnenberg HG: **Histone modification patterns associated with the human X chromosome**. *EMBO Rep* 2006, **7**(6):628-634.

9. Fraga MF, Ballestar E, Villar-Garea A, Boix-Chornet M, Espada J, Schotta G, Bonaldi T, Haydon C, Ropero S, Petrie K *et al*: **Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer**. *Nat Genet* 2005, **37**(4):391-400.
10. Esteller M: **Cancer epigenomics: DNA methylomes and histone-modification maps**. *Nat Rev Genet* 2007, **8**(4):286-298.
11. Berger SL: **The complex language of chromatin regulation during transcription**. *Nature* 2007, **447**(7143):407-412.
12. Khan AU, Krishnamurthy S: **Histone modifications as key regulators of transcription**. *Front Biosci* 2005, **10**:866-872.
13. Schubeler D, MacAlpine DM, Scalzo D, Wirbelauer C, Kooperberg C, van Leeuwen F, Gottschling DE, O'Neill LP, Turner BM, Delrow J *et al*: **The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote**. *Genes Dev* 2004, **18**(11):1263-1271.
14. Bernstein BE, Kamal M, Lindblad-Toh K, Bekiranov S, Bailey DK, Huebert DJ, McMahon S, Karlsson EK, Kulbokas EJ, 3rd, Gingeras TR *et al*: **Genomic maps and comparative analysis of histone modifications in human and mouse**. *Cell* 2005, **120**(2):169-181.
15. Liu CL, Kaplan T, Kim M, Buratowski S, Schreiber SL, Friedman N, Rando OJ: **Single-nucleosome mapping of histone modifications in S. cerevisiae**. *PLoS Biol* 2005, **3**(10):e328.
16. Millar CB, Grunstein M: **Genome-wide patterns of histone modifications in yeast**. *Nat Rev Mol Cell Biol* 2006, **7**(9):657-666.
17. Kurdistani SK, Tavazoie S, Grunstein M: **Mapping global histone acetylation patterns to gene expression**. *Cell* 2004, **117**(6):721-733.
18. Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh TY, Peng W, Zhang MQ *et al*: **Combinatorial patterns of histone acetylations and methylations in the human genome**. *Nat Genet* 2008, **40**(7):897-903.
19. Ercan S, Giresi PG, Whittle CM, Zhang X, Green RD, Lieb JD: **X chromosome repression by localization of the C. elegans dosage compensation machinery to sites of transcription initiation**. *Nat Genet* 2007, **39**(3):403-408.
20. Ercan S, Dick LL, Lieb JD: **The C. elegans dosage compensation complex propagates dynamically and independently of X chromosome sequence**. *Curr Biol* 2009, **19**(21):1777-1787.
21. Cairns BR: **The logic of chromatin architecture and remodelling at promoters**. *Nature* 2009, **461**(7261):193-198.
22. Gelato KA, Fischle W: **Role of histone modifications in defining chromatin structure and function**. *Biol Chem* 2008, **389**(4):353-363.
23. Saha A, Wittmeyer J, Cairns BR: **Chromatin remodelling: the industrial revolution of DNA around histones**. *Nat Rev Mol Cell Biol* 2006, **7**(6):437-447.
24. Strahl BD, Allis CD: **The language of covalent histone modifications**. *Nature* 2000, **403**(6765):41-45.
25. Jenuwein T, Allis CD: **Translating the histone code**. *Science* 2001, **293**(5532):1074-1080.
26. Turner BM: **Defining an epigenetic code**. *Nat Cell Biol* 2007, **9**(1):2-6.

27. Suganuma T, Workman JL: **Crosstalk among Histone Modifications**. *Cell* 2008, **135**(4):604-607.
28. Dion MF, Altschuler SJ, Wu LF, Rando OJ: **Genomic characterization reveals a simple histone H4 acetylation code**. *Proc Natl Acad Sci U S A* 2005, **102**(15):5501-5506.
29. van Leeuwen F, van Steensel B: **Histone modifications: from genome-wide maps to functional insights**. *Genome Biol* 2005, **6**(6):113.
30. Celniker SE, Dillon LA, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai EC, Lieb JD, MacAlpine DM *et al*: **Unlocking the secrets of the genome**. *Nature* 2009, **459**(7249):927-930.
31. Pillai S, Chellappan SP: **ChIP on chip assays: genome-wide analysis of transcription factor binding and histone modifications**. *Methods Mol Biol* 2009, **523**:341-366.
32. Schones DE, Zhao K: **Genome-wide approaches to studying chromatin modifications**. *Nat Rev Genet* 2008, **9**(3):179-191.
33. Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC, Suzuki M, Kawai J *et al*: **Antisense transcription in the mammalian transcriptome**. *Science* 2005, **309**(5740):1564-1566.
34. Baugh LR, Demodena J, Sternberg PW: **RNA Pol II accumulates at promoters of growth genes during developmental arrest**. *Science* 2009, **324**(5923):92-94.
35. Core LJ, Waterfall JJ, Lis JT: **Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters**. *Science* 2008, **322**(5909):1845-1848.
36. Seila AC, Calabrese JM, Levine SS, Yeo GW, Rahl PB, Flynn RA, Young RA, Sharp PA: **Divergent transcription from active promoters**. *Science* 2008, **322**(5909):1849-1851.
37. Bender LB, Suh J, Carroll CR, Fong Y, Fingerman IM, Briggs SD, Cao R, Zhang Y, Reinke V, Strome S: **MES-4: an autosome-associated histone methyltransferase that participates in silencing the X chromosomes in the C. elegans germ line**. *Development* 2006, **133**(19):3907-3917.
38. Takasaki T, Liu Z, Habara Y, Nishiwaki K, Nakayama J, Inoue K, Sakamoto H, Strome S: **MRG-1, an autosome-associated protein, silences X-linked genes and protects germline immortality in Caenorhabditis elegans**. *Development* 2007, **134**(4):757-767.
39. Blumenthal T, Evans D, Link CD, Guffanti A, Lawson D, Thierry-Mieg J, Thierry-Mieg D, Chiu WL, Duke K, Kiraly M *et al*: **A global analysis of Caenorhabditis elegans operons**. *Nature* 2002, **417**(6891):851-854.
40. Reinke V: **Functional exploration of the C. elegans genome using DNA microarrays**. *Nat Genet* 2002, **32 Suppl**:541-546.
41. Blumenthal T, Gleason KS: **Caenorhabditis elegans operons: form and function**. *Nat Rev Genet* 2003, **4**(2):112-120.
42. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ: **miRBase: tools for microRNA genomics**. *Nucleic Acids Res* 2008, **36**(Database issue):D154-158.
43. Kato M, de Lencastre A, Pincus Z, Slack FJ: **Dynamic expression of small non-coding RNAs, including novel microRNAs and piRNAs/21U-RNAs, during Caenorhabditis elegans development**. *Genome Biol* 2009, **10**(5):R54.

44. Martinez NJ, Ow MC, Barrasa MI, Hammell M, Sequerra R, Doucette-Stamm L, Roth FP, Ambros VR, Walhout AJ: **A C. elegans genome-scale microRNA network contains composite feedback motifs with high flux capacity**. *Genes Dev* 2008, **22**(18):2535-2549.
45. Barski A, Jothi R, Cuddapah S, Cui K, Roh TY, Schones DE, Zhao K: **Chromatin poises miRNA- and protein-coding genes for expression**. *Genome Res* 2009, **19**(10):1742-1751.
46. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K: **High-resolution profiling of histone methylations in the human genome**. *Cell* 2007, **129**(4):823-837.
47. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP *et al*: **Genome-wide maps of chromatin state in pluripotent and lineage-committed cells**. *Nature* 2007, **448**(7153):553-560.
48. Karlic R, Chung HR, Lasserre J, Vlahovicek K, Vingron M: **Histone modification levels are predictive for gene expression**. *Proc Natl Acad Sci U S A* 2010, **107**(7):2926-2931.
49. Kouzarides T: **Chromatin modifications and their function**. *Cell* 2007, **128**(4):693-705.
50. Sims RJ, 3rd, Reinberg D: **Is there a code embedded in proteins that is based on post-translational modifications?** *Nat Rev Mol Cell Biol* 2008, **9**(10):815-820.
51. Schreiber SL, Bernstein BE: **Signaling network model of chromatin**. *Cell* 2002, **111**(6):771-778.
52. Ng HH, Robert F, Young RA, Struhl K: **Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity**. *Mol Cell* 2003, **11**(3):709-719.
53. Li J, Moazed D, Gygi SP: **Association of the histone methyltransferase Set2 with RNA polymerase II plays a role in transcription elongation**. *J Biol Chem* 2002, **277**(51):49383-49388.
54. Fischer JJ, Toedling J, Krueger T, Schueler M, Huber W, Sperling S: **Combinatorial effects of four histone modifications in transcription and differentiation**. *Genomics* 2008, **91**(1):41-51.
55. Fuchs SM, Laribee RN, Strahl BD: **Protein modifications in transcription elongation**. *Biochim Biophys Acta* 2009, **1789**(1):26-36.
56. Chambeyron S, Bickmore WA: **Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription**. *Genes Dev* 2004, **18**(10):1119-1130.
57. Harris TW, Antoshechkin I, Bieri T, Blasiar D, Chan J, Chen WJ, De La Cruz N, Davis P, Duesbury M, Fang R *et al*: **WormBase: a comprehensive resource for nematode research**. *Nucleic Acids Res* 2010, **38**(Database issue):D463-467.
58. Cristianini N, Shawe-Taylor J: **An Introduction to Support Vector Machines and other kernel-based learning methods**: Cambridge University Press; 2000.
59. Wang Y, Liu CL, Storey JD, Tibshirani RJ, Herschlag D, Brown PO: **Precision and functional specificity in mRNA decay**. *Proc Natl Acad Sci U S A* 2002, **99**(9):5860-5865.

60. Pokholok DK, Harbison CT, Levine S, Cole M, Hannett NM, Lee TI, Bell GW, Walker K, Rolfe PA, Herbolsheimer E *et al*: **Genome-wide map of nucleosome acetylation and methylation in yeast**. *Cell* 2005, **122**(4):517-527.
61. Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G *et al*: **Stem cell transcriptome profiling via massive-scale mRNA sequencing**. *Nat Methods* 2008, **5**(7):613-619.
62. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB *et al*: **Genome-scale DNA methylation maps of pluripotent and differentiated cells**. *Nature* 2008, **454**(7205):766-770.
63. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq**. *Nat Methods* 2008, **5**(7):621-628.

Figure Legends

Figure 1: Schematic diagram of our data binning and supervised analysis. (A) DNA regions around the transcription start site (TSS) and transcription terminal site (TTS) of each transcript were separated into 160 bins of 100 bp in size. Average signal of each chromatin feature was calculated for all transcripts, resulting in a predictor matrix for each bin. These predictor matrices were used to predict expression of transcripts by support vector machine (SVM) or support vector regression (SVR) models. The genome-wide data for chromatin features and gene expression were generated by the modENCODE project using ChIP-chip/ChIP-seq and RNA-seq experiments, respectively. (B) A summary of datasets used in our analysis.

Figure 2: Signal pattern (A) and correlation pattern (B) of each chromatin feature in the 160 bins around the TSS and TTS (from 4kb upstream to 4kb downstream) of worm transcripts at the EEMB stage. In (A), signal of each chromatin feature for each bin is averaged across all transcripts. In (B), the Spearman correlation coefficient of each chromatin feature with gene expression levels was calculated for each bin. Ab1 and Ab2 represent experimental results using different antibodies for a chromatin feature. DNA region from 2kb upstream of the TSS to 2kb downstream of the TTS is shown in the rectangle.

Figure 3: Hierarchical clustering using either chromatin feature profiles (A-C) or bin profiles (D-E) discriminates highly and lowly expressed genes. (A): Hierarchical clustering of 16 chromatin features in Bin #1 (0-100 nt upstream of TSS). The resulting tree is split at the top branch, which divides genes into two clusters, Cluster H and Cluster L, as labeled. (B): Distributions of expression levels of genes in Cluster H (red) and Cluster L (green). Expression levels are significantly different between the two clusters according to t-test (P=3E-202). Expression levels were measured by RNA-seq (see Methods). (C): T-scores for the differential expression of the top two gene clusters based on hierarchical clustering of chromatin features in each of the 160 bins. For each bin, hierarchical clustering was performed to separate genes into two clusters. Expression levels between the two clusters were compared and a t-score calculated to measure the capability of the bin to discriminate between genes with high and low expression levels. (D): Hierarchical clustering of the genes based on the signal profiles of H3K79me2 across the 160 bins. The resulting tree is also split at the top branch, leading to two gene clusters. (E): Distributions of expression levels of genes in the two clusters in (D). The expression levels are significantly different according to t-test (P=4E-93). (F): T-scores for the differential expression of the two gene clusters based on hierarchical clustering of bin profiles for each individual chromatin feature. Cyan and blue colors indicate a significant positive and negative correlation between a chromatin feature and gene expression levels, respectively. Black color indicates that a chromatin feature could not significantly discriminate between genes with high and low expression levels. To visualizing the clustering, 2,000 randomly selected genes are shown. The data for gene expression levels and chromatin features are from EEMB stage.

Figure 4: Prediction power of the supervised models. (A): ROC curves for 5 different bins based on the results of the SVM classification models. (B): Predicted versus experimentally measured expression levels. The SVR regression model was applied to Bin #1 for predicting gene expression levels. (PCC: Pearson correlation coefficient) (C): The prediction accuracy of SVM classification models for all the 160 bins. For each bin, we constructed an SVM classification model and summarized its accuracy using the AUC score. The AUC scores were calculated based on cross-validation repeated 100 times for each bin. The red curve shows the average AUC scores (mean of 100 repeats) of the bins and the blue bars indicate their standard deviations. The positions of TSS and TTS are marked by dotted lines.

Figure 5: Prediction power of the SVM models using the signals from different subsets of chromatin features in the 100 nt around TSS (bin 1). The results are based on cross-validation with 100 trials. (A): ALL- all 21 chromatin features; H3- the two H3 features; HIS- the 11

chromatin modification features; XIF- the seven binding profile features for X-inactivation factors; POLII- the binding profile feature for RNA polymerase II. (B): HIS- the 11 chromatin modification features; H3K79ME- H3K79me1, H3K79me2 and H3K79me3; H3K9ME- H3K9me2, H3K9me3(Ab1) and H3K9me3(Ab2); H3K36ME- H3K36me2(Ab1), H3K36me2(Ab2) and H3K36me3; H3K4ME-H3K4me3 and H3K4me3.

Figure 6: Co-regulation of transcription by pairs of histone modifications. (A): Categorization of genes into four groups based on signals of H3K4me3 and H3K36me3: HH (magenta), HL (green), LH (cyan) and LL (blue). The signals of histone marks H3K36me3, H3K4me3 exhibit a bimodal feature. Signals are thus classified into H and L by a Gaussian mixture model. The distributions of expression levels of the four gene groups are shown on the right. (B): Same as (A), based on signals of H3K9me3 and H3K79me3. Same as above, the signal of H3K79me3 is again classified by a Gaussian mixture model. The signals of H3K9me3 do not display a bimodal feature, signals are classified into H and L based on whether the value is higher than or lower than the median.

Figure 7: Developmental stage specificity of the chromatin model. The EEMB model was constructed using the chromatin features and gene expression data both at the EEMB stage. The model was then used to predict gene expression levels at EEMB and 5 other developmental stages: L1, L2, L3, L4 and Adult. ROC curves are plotted based on the results of 100 trials of cross-validation. For each trial, the dataset was randomly separated into two halves: one half as training data and the other as testing data to estimate the accuracy of the model. The values in parentheses are AUC scores.

Figure 8: Correlation patterns of H3K4me3 and H3K79me3 in the 160 bins around the TSS and TTS (from 4kb upstream to 4kb downstream) with the expression levels of the first, second and last genes of 881 *C. elegans* operons.

Figure 9: Prediction of expression levels of microRNAs at EEMB. (A): Predicted expression levels of the experimentally measured highly and lowly expressed microRNAs based on small RNA-seq results. Expression levels of microRNAs at EEMB stage were predicted using an SVR regression model trained on data for protein-coding genes at the same stage. (B): Predicted versus experimentally measured expression levels of microRNAs at the EEMB stage. R is the Pearson correlation coefficient.

Figure 10: Prediction accuracy of the chromatin model in four other species. Expression levels of genes are predicted using the SVR method. In yeast, average signals of chromatin features from the translation start site to 500bp upstream were used as predictors (A); in the other species, signals of chromatin features within the bin at the TSS (bin#1) were used as predictors (B-D). E4-8h: embryonic stage at 4-8h. ESC: embryonic stem cell.

Additional Files

Additional File 1
Title: Signal patterns of Pol II around TSS and TTS regions (from -4kb to 4kb) at different developmental stages.
Format: PDF file
Description: At each stage, the signals were normalized by subtracting the average and then divided by the standard deviation of the signals over all the 160 bins. The location of TSS and TTS are marked as dotted lines.

Additional File 2
Title: Correlation patterns of chromatin features with gene expression at the EEMB stage based on long transcript gene only.
Format: PDF file
Description: Only genes longer than 8kb were used for correlation computations so that there is no overlap between the TSS and TTS bins.

Additional File 3
Title: Correlation patterns of chromatin features with gene expression at the EEMB stage based on transcripts that are far away from any other transcripts.
Format: PDF file
Description: Only the transcripts that are at least 4kb away from any other transcripts were used for correlation computations so that there is no overlap between bins of nearby transcripts.

Additional File 4
Title: Correlation patterns of chromatin features with gene expression at the L3 stage. Correlation was calculated based on long transcripts (>8kb).
Format: PDF file
Description:

Additional File 5
Title: Correlation patterns of chromatin features with gene expression at the EEMB stage based on single-transcript genes only.
Format: PDF file
Description:

Additional File 6
Title: Prediction of gene expression using chromatin features in all the 40 bins around TSS (from -2kb to 2kb).
Format: PDF file
Description: A: ROC curve of the SVM classification model. B: Predicted expression levels versus actual expression levels measured by RNA-seq experiment. PCC is the Pearson correlation coefficient.

Additional File 7
Title: Interaction between all possible pairs of histone modifications.
Format: EXCEL file
Description: Interaction between all possible pairs of histone modification as indicated by linear model in Bin #1. For each pair, both the results of linear models with the interaction terms (Interaction models) and without the interaction terms (Singleton models) are shown.

Additional File 8
Title: The significant interactions between chromatin features based on a linear model.
Format: EXCEL file
Description: The significant interactions between chromatin features based on a linear model with 12 different chromatin features and their pairwise interaction terms.

Additional File 9
Title: Mutual information between expression and pairwise histone modification signals.
Format: PDF file
Description: For each pair of histone modifications (denoted as H1, H2), the heat map shows the normalized mutual information I(E, H1 AND H2)/max(I(E,H1),I(E,H2)). For pairs such as H3K4me2 and K4K36me3, the combination of two features gives a higher predictive power than the two individual features.

Additional File 10
Title: Interactions among chromatin features and expression.
Format: PDF file
Description: (A) Node colors indicate the correlation of the corresponding features with gene expression. Edge colors indicate the correlation between the two connected features. Only interactions with a strong correlation (|PCC| > 0.3) are shown. (B) The directional relationships inferred from Bayesian network analysis. Arrow sizes indicate the confidence scores of the directed edges. Only interactions with a confidence score (combined for both directions) of at least 80% are shown.

Additional File 11
Title: Supplementary documents about the Bayesian network analysis etc.
Format: PDF file
Description: The file contains additional information about the Bayesian network analysis.

Additional File 12
Title: Correlation patterns of chromatin features in 40 bins around TSS and TTS of the first and the second genes in 881 worm operons.
Format: PDF file
Description: Correlation patterns of chromatin features in 40 bins around TSS and TTS (from -2kb to 2kb) of the first and the second genes in 881 worm operons.

Additional File 13
Title: Predicted expression levels of microRNAs at stage L3.
Format: PDF file
Description: MicroRNAs are divided into High (red) and Low (green) groups based on their measured expression levels in small RNA-seq experiments.

Additional File 14
Title: Stage specificity of chromatin models for microRNAs expression predictions.
Format: PDF file
Description: The chromatin model was trained using the chromatin and expression data of protein-coding genes at the EEMB stage. The model was then used to predict microRNA expression levels at six stages. R indicates the Pearson correlation coefficient between the predicted expression levels and the actual expression levels from RNA-seq experiments.

Additional File 15
Title: Gene Expression Omnibus accession ID of data sets.
Format: EXCEL file
Description: The file contains the GEO ID of data sets used in this work.