Recent epigenomic studies have identified significant differences between developmental stages and cell types. While the importance of epigenetic regulation has been increasingly recognized, it remains unclear how the global epigenetic patterns are established and maintained. Here I review a number of recent studies with the emphasis on the role of the genomic sequence in shaping the epigenetic landscape. These studies strongly suggest that the sequence information is important not just for controlling target specificity but for orchestrating the diversity of epigenetic patterns among different cell types. The epigenome is maintained by the complex network of a large number of interactions. Integrative approaches are needed to gain insights into these networks.

INTRODUCTION

The genome provides a blueprint for gene regulation, but this blueprint is interpreted differently in different cell types. A partial explanation for these differences is the fact that the DNA sequences are packaged into chromatin, and that only a small portion of the genome is accessible in any cell type. DNA accessibility in a cell is highly controlled by the combined effects of multiple epigenetic marks including nucleosome positions, histone modifications, and DNA methylation.1

The basic repeating unit of chromatin is the nucleosome, which consists of 147 base pairs of DNA wrapped around a core histone octamer in nearly two times.2 The nucleosome is a formidable barrier for transcription factors (TF) binding, thereby serving as a repressor for transcription. The N-terminal histone tails can be covalently modified at numerous locations in different ways. The combinatorial pattern not only serves as a landmark for open or closed chromatin but recruit specific regulatory proteins.3,4 Epigenetic information can also be carried by the genomic DNA itself, where certain cytosine nucleotides are methylated at the C-5 position.5 While promoter DNA methylation has been long recognized to be associated with gene silencing, genome-wide studies have identified new functions associated with coding region methylation.6,7

During the past decade a large amount of epigenomic data has been generated (recently reviewed in Refs 8–10), providing high-resolution maps of the entire epigenome which can be compared to gain functional insights. Significant differences have been identified between cells under different growth conditions or at developmental stages. What is more clinically relevant is the identification of global epigenetic aberrations in cancer and a number of other diseases,11,12 suggesting dysfunctional epigenetic regulation may be an important step to these diseases.

Despite these exciting discoveries, we are still far from understanding the mechanisms that regulate the genome-wide epigenetic patterns. A number of fundamental questions remain unresolved. How are the chromatin regulators and DNA methyltransferases recruited to the chromatin? How is the global epigenetic pattern stably maintained after initial establishment? What are the factors that cause epigenetic aberrations during the development of a certain disease? Is it possible to restore normal epigenetic patterns by clinical intervention? Answers to these questions are needed both for understanding the function of epigenetic mechanisms and for developing effective approaches for disease treatment.

It has been increasingly recognized that epigenomic patterns are regulated by the complex, dynamic interactions among multiple classes of factors, including chromatin modifiers, DNA binding proteins,
noncoding RNAs (ncRNA), and signaling molecules. In this article, I will focus on the role of DNA sequence in mediating these interactions. This is a central question that has been extensively investigated yet remains incompletely understood. More comprehensive discussions can be found in several recent reviews.\textsuperscript{13–16}

**COMPUTATION METHODS FOR PREDICTIONS OF EPIGENOMIC PATTERNS**

Sequence analysis has traditionally been used to predict TF binding sites, which often target specific short DNA sequence patterns called motifs (reviewed in Ref 17). However, most chromatin regulators either do not or weakly interact with DNA. While interactions with certain TFs may play an important role in recruitment, such interactions often involve many TFs. As a result, an epigenetic pattern is often associated with many sequence features each only making a minor contribution. A recent trend in computational biology is to develop effective predictive models by integrating multiple weak features, as described below.

A commonly used approach is supervised learning.\textsuperscript{18} The availability of genome-wide data makes it possible to select sample sequences from either positive or negative regions. In order to build a predictive model, it is important to convert these sequences to numerical values. If the relevant sequence features are known, their occurring frequencies can be used as numerical predictors. However, in most cases discriminative sequence features are not known \textit{a priori}. To address this problem, various classes of sequence features have been explored, including TF motifs,\textsuperscript{19–21} word counts,\textsuperscript{22} repetitive elements,\textsuperscript{19,21,23} DNA structural parameters,\textsuperscript{24} and periodic patterns.\textsuperscript{25} Table 1 shows an incomplete list of DNA sequence features and their associated factors.

<table>
<thead>
<tr>
<th>Sequence Features</th>
<th>Recruiting Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>CpG</td>
<td>CXXC domain containing factors; transcription factors; core histone; methylation sensitive binding factors</td>
</tr>
<tr>
<td>Poly A:T</td>
<td>Core histone</td>
</tr>
<tr>
<td>TF motifs</td>
<td>Transcription factors</td>
</tr>
<tr>
<td>Repetitive elements</td>
<td>RNAi machineries</td>
</tr>
<tr>
<td>Boundary elements</td>
<td>Insulator proteins; transcription factors</td>
</tr>
<tr>
<td>Uncharacterized</td>
<td>Noncoding RNA</td>
</tr>
</tbody>
</table>

Because of the large number of potentially useful sequence features (it is not uncommon to have more than 100 features), one important concern is overfitting—failure of a complex model that is trained based on a small sample to generate useful predictions. This risk can be reduced by using variable selection techniques. This can be done either in the traditional linear regression framework, by using principal component analysis, stepwise regression, penalized regression,\textsuperscript{26} or through machine learning approaches such as decision trees\textsuperscript{27} and support vector machines.\textsuperscript{28} Finally, prediction accuracy can be improved by ensemble-based prediction models such as boosting\textsuperscript{29} and Bayesian additive regression trees (BART).\textsuperscript{30} These methods have been applied to prediction of nucleosome positioning,\textsuperscript{22,24,25} histone modifications,\textsuperscript{20,31,32} and DNA methylation.\textsuperscript{19–21,33} The performance of different methods in predicting protein–DNA interactions has been compared, with the results in favor of boosting and BART.\textsuperscript{34}

An alternative approach is to estimate the position specific sequence pattern directly as an extension of traditional motif analysis methods.\textsuperscript{35} In this approach, the input sequences are first aligned, and then the probability of observing a certain nucleotide at a specific position is empirically estimated and further normalized. This approach has been applied to predict nucleosome positioning. Accurate predictions can be achieved when complemented by word frequency features.\textsuperscript{36,37}

Another approach for nucleosome positioning prediction is based on biophysical properties.\textsuperscript{38–40} Viewing DNA as a flexible polymer, the sequence dependent, free energy associated with a specific geometric configuration can be calculated. As the nucleosome crystal structure has been experimentally identified,\textsuperscript{41} the most favorable sequence composition can be identified by minimizing the free energy.\textsuperscript{38–40} While these models do not require genome-wide data for model training, the prediction accuracy is surprisingly competitive. However, the utility of this approach to other epigenetic marks is limited because of the lack of high-resolution protein structure information.

The basic output of each method is a propensity score for an input sequence fragment; this can be extended to predict genome-wide epigenetic patterns by scanning with sliding windows. Additional constraints, such as the steric hindrance effect of neighboring nucleosomes, can be further incorporated, for example, by using a hidden Markov model.\textsuperscript{35,42} An extension of this approach has been applied to predict multiple factor binding profiles.\textsuperscript{33,44}
QUANTITATIVE EVALUATION OF THE ROLE OF DNA SEQUENCES IN REGULATING EPIGENOMIC PATTERNS

The computational methods described above have been applied to quantitatively evaluate the association between DNA and epigenomic patterns. As discussed below, while each epigenetic mark involves a distinct set of sequence features and recruiting factors, the underlying principle is strikingly similar.

Nucleosome Positioning

The nucleosome is the fundamental unit for packaging DNA. Its highly conserved three dimensional structure was identified by Richmond and colleagues, showing more than 120 direct histone–DNA interactions unevenly distributed across the nucleosome surface. Each individual interaction is rather weak, allowing the flexibility to package almost the entire genome. A long-standing question is to what extent the genome-wide nucleosome positioning is dictated by such variations.

The development of genomic technologies and computational methodologies has greatly advanced our understanding of the role of DNA sequence in global regulation of nucleosome positioning. The best studied model system is Saccharomyces cerevisiae. Here high-resolution, genome-wide mapping of nucleosome positions has been obtained by crosslinking nucleosome with DNA, followed by MNase digestion and microarray hybridization or high-throughput sequencing. A number of computational methods have been developed to predict nucleosome positions from the genomic sequence.

While the prediction accuracy is significantly better than random guess, it is highly variable across the genome (see Ref 50 for a review). While the nucleosome-free region (NFR) and +1 nucleosome position at the 5′ end of genes can be well predicted, the accuracy in other regions is poorer. One significant limitation is that while the in vivo nucleosome positions display regular phasing surrounding NFRs, predicted nucleosome positions do not have this pattern.

To test whether the DNA sequences indeed play a causal role, several groups have carried out experimental validations by deleting certain sequences that are predicted to be associated with nucleosome positions followed by comparison of the nucleosome positions in wild-type and mutants. The nucleosome positions in the mutants are indeed altered, but the degree of alteration gradually increases with respect to the deletion size, suggesting that nucleosome positions are mediated by the cumulative effect of many weak histone–DNA interactions.

Two independent studies have used a similar approach to investigate the direct role of DNA sequences in nucleosome positioning. In both studies, the yeast genome was extracted from the nucleus and reassembled into nucleosomes in a cell-free media. The in vitro assembled nucleosome positioning pattern was then detected by using next generation sequencing. Both studies have found that the in vitro and in vivo nucleosome occupancy patterns are highly correlated; notably, the DNA sequence of most yeast promoters intrinsically disfavor nucleosome assembly. On the other hand, the center positions of the nucleosomes vary significantly between the two conditions. The nucleosomes assembled in vitro are more delocalized, suggesting that the DNA sequence plays an important but insufficient role for genome-wide nucleosome positioning.

One striking feature of in vivo nucleosome positioning pattern is the regular phasing of nucleosomes around NFR, but such a pattern cannot be reproduced by in vitro nucleosome assembly. This difference can be explained by a simple model, called statistical positioning, proposed by Kornberg and Stryer more than twenty years ago. This model assumes that nucleosome positions are bounded by a fixed barrier but distributed at random elsewhere. If averaged over all possible configurations, similar to averaging over a large cell population, the resulting pattern appears to be regular. Despite the elegance of the above model, recent experimental studies have uncovered important additional complexity. In one study, the investigators reconstituted nucleosomes in a whole-cell-extract media supplemented by ATP and were able to recapitulate the in vitro-like nucleosome positioning pattern. Notably, the spacing between neighboring nucleosomes is not strongly dependent upon the local nucleosome density, as predicted by the statistical positioning model. These findings strongly suggest that in vivo nucleosome positions are not simply formed by statistical positioning but result from active packing by ATP-dependent remodeling complexes.

There is a great interest in to what extent the DNA sequence is conserved across different species. Genome-wide nucleosome positions have been mapped in a number of organisms, including Schizosaccharomyces pombe, Caenorhabditis elegans, Drosophila melanogaster, Arabidopsis thaliana, and human. All these species share a similar pattern, that is, most promoters contain an NFR flanked by regular nucleosome arrays on each side. This is striking, considering that the
promoter sequences have diverged significantly during evolution. For example, while the GC-content is typically low for a yeast promoter, it is significantly higher in mammalian promoters, which often harbor a CpG island. Indeed, computational studies have predicted human promoters to be encoded for high nucleosome occupancy. One explanation of this counter-intuitive result is that most human genes are tissue-specific therefore are silenced by default and overridden by transcriptional regulators in a cell type specific manner. This view is supported by in vitro nucleosome positioning data, which show no depletion of nucleosomes at CpG islands. Similar to yeast, the in vitro data do not show regular phasing. In addition, the CpG islands seem to affect the function of the SWI/SNF remodeling complexes, which are required for induction of only those genes harboring CpG island promoters.

**Histone Modification**

Investigations of the regulatory mechanisms for histone modification patterns are complicated by the fact that there are a large number of modifications each having its own global distributions. Furthermore, genome-wide studies have been limited to those histone marks for which high-quality antibodies are available. An important role of the DNA sequence has been recognized, but it is also possible that certain modifications are regulated mainly through sequence independent mechanisms.

There are three histone marks whose recruiting mechanisms have been systematically investigated: H3K4me3, H3K27me3, and H3K9me3. While H3K4me3 is an active mark that is typically present at the transcription start sites of active genes, both H3K27me3 and H3K9me3 marks are associated with gene silencing. Despite their opposite functions, H3K4me3 and H3K27me3 colocalize at certain genomic regions called bivalent domains, which play an important role in development by marking genes that are poised for activation during cell differentiation. The H3K9me3 mark is a typically associated with heterochromatin but can also be found at certain euchromatic regions.

The genome-wide distribution of H3K4me3 can be predicted with high accuracy from the DNA sequence in different cell types. There is a high degree of overlap between H3K4me3 targets and CpG islands, which is at least in part due to selective binding of the Setd1/MLL complex member Cfp1, which is a CXXC domain containing protein, to non-methylated CpG. Depletion of Cfp1 substantially reduces H3K4me3 recruitment at the CpG islands. Furthermore, H3K4me3 recruitment in ES cells is also mediated by interaction between Wdr5, another component in Setd1/MLL complex and Oct4.

While the H3K4me3 distribution is mainly characterized by sharp peaks (<1 kb), the distribution of H3K27me3 is more complex. It not only has peaks at selected promoters but also spans over broad domains. The global distribution of H3K27me3 highly colocalizes with Polycomb complexes, which contain the catalytic unit for H3K27 methylation. The core PcG complexes are highly evolutionarily conserved. Sequence-based models have been developed to predict PcG target genes in mammalian ES cells, and many of target genes can be predicted based on the CpG density alone. An intriguing hypothesis is that PcG complexes target high CpG density regions by default. A direct relationship between high CpG density and PcG recruitment has been experimentally verified by transgenic assays. On the other hand, it is likely that the recruitment can be further modulated by TFs and ncRNAs. Incorporating TF motifs, such as Myc and E2F, in a computational model can substantially improve the accuracy for PcG target predictions.

An important question is what DNA elements are required for PcG recruitment, and this question cannot be simply answered by genome-wide location studies. Such DNA elements are referred to as Polycomb response elements (PRE). The PREs are well-characterized in Drosophila, containing binding sites for a number of TFs, such as Pho and the GAGA factor. Sequence-based computational models can predict PREs locations with high accuracy. Importantly, in one study the investigators followed up by experimentally validation using transgenic assays and found that 29 out of 43 predicted PREs can be experimentally validated. In contrast, only few PREs have been experimentally identified in mammals. There is growing evidence that the recruiting mechanism for mammals is fundamentally different; in particular, ncRNAs seem to play an important role in genome-wide recruitment of PcG complexes.

Genome-wide analysis of the H3K9me3 is more difficult compared to the two marks discussed above, in part due to its strong association with regions which cannot be uniquely mapped to the reference genome. The distribution over mappable regions is much more diffusive compared to the two marks discussed above. The prediction accuracy for the H3K9me3 distribution is also lower, suggesting that the role of DNA sequence information is weaker. However, these observations cannot exclude the possibility that the DNA sequence still plays an important role at the recruitment step, and once recruited, the histone mark spreads along the chromosome.
in a sequence independent manner. Supporting this hypothesis, it has been shown that the RNA transcribed from repetitive DNA elements can promote heterochromatin formation through an RNAi dependent pathway.\(^\text{80,81}\) Deletion of these repetitive elements can abandon both heterochromatin formation and H3K9me3 recruitment.\(^\text{80}\) While these experiments were done in S. pombe, such a mechanism seems to be conserved among eukaryotes.\(^\text{82}\) In mammals, H3K9me3 appears to be more enriched in satellites than other classes of repetitive sequences.\(^\text{83}\)

In addition to the three histone marks discussed above, the DNA sequence is likely to play an important role in regulating chromatin states in a general sense. Recently, computational models have been developed to identify chromatin states from combinatorial histone modification patterns.\(^\text{84-87}\) One study has investigated the link between chromatin states and DNA sequence patterns and found that a number of TF motifs are enriched with the enhancer chromatin state.\(^\text{88}\) Furthermore, computational methods have been developed to predict p300 binding sites from DNA sequence patterns with good accuracy.\(^\text{89}\) Interestingly, the role of DNA sequence for regulating the pattern of H3K4me1, a mark for enhancer regions, seems to be highly cell-type specific, and is strongest in ES cells.\(^\text{90}\)

**DNA Methylation**

In vertebrate animals, DNA methylation almost exclusively occurs in the context of CpG dinucleotides.\(^\text{8}\) Genome-wide DNA methylation level is highly associated with the CpG density and can be predicted with high-accuracy.\(^\text{19,20}\) While CpG islands are normally unmethylated, a subset may become hypermethylated in cancer, which may silence tumor suppressor genes thereby playing an important role in tumorigenesis.\(^\text{11}\) Several groups have developed algorithms to predict hypermethylated CpG islands with moderate accuracy despite being statistically significant.\(^\text{19-21,23}\) In addition to CpG density, several TF motifs, including Sp1, NRF1, and YY1, have also been found to be methylation resistant, and motifs such as GAGA have been found to be methylation prone.\(^\text{91}\) Most of these motifs are also associated with PcG targets,\(^\text{32,70}\) which is not surprising as hypermethylated genes in cancer significantly overlap with PcG targets in ES cells.\(^\text{92}\) In general, DNA methylation and chromatin state patterns are closely related, but a detailed discussion is beyond the scope of this review. Also, repetitive sequences, such as the Alu elements, have been found to be enriched with DNA methylation.\(^\text{19,93}\) In addition, a motif sequence associated with the G-quadruplex structure has been shown to be associated with hypomethylation.\(^\text{94}\) In this study, the investigators also found that hypomethylated G-quadruplex sequence motif is enriched at the DNA breakpoints in many cancer types,\(^\text{94}\) suggesting a link between epigenetic pattern and genome stability.

With the rapid technological development, recent studies have been able to profile DNA methylation genome-wide.\(^\text{6,95-97}\) As a result, a more dynamic picture has emerged: hyper- and hypo-methylation are no longer viewed as isolated events but likely the results of dynamic boundary shift near CpG island shores.\(^\text{95,96}\) While it is clear that such transitions are mediated by the CpG density, it will be interesting to investigate whether there are additional DNA elements that are involved in the regulation.

A causal role of DNA elements in regulating DNA methylation pattern has been recently experimentally validated by using a transgenic approach.\(^\text{98}\) These authors inserted selected promoter DNA sequences into the β-globin locus in mouse embryonic stem cells and measured the DNA methylation level by using bisulfite sequencing. They found that the DNA methylation pattern at the inserted sites is quite similar to the endogenous sites, suggesting that the DNA sequence is sufficient for maintenance of DNA methylation pattern. Furthermore, they found that CpG density alone is not sufficient for maintenance of DNA methylation pattern, but it also requires combination of TF motifs.

**INSIGHTS INTO RECRUITMENT FACTORS**

Computational sequence analyses are useful not only to predict regions targeted by a specific epigenetic factor but also to generate hypotheses regarding the recruiting factors, which can be validated in the laboratory. Here we briefly discuss the role of three general classes of recruiting factors.

**Transcription Factors**

As chromatin regulators often cannot directly bind DNA, it has been long speculated that their recruitment may be mediated by interactions with TFs. The sequence analysis results discussed above strongly support this hypothesis. In addition, recent proteomic studies have identified extensive interactions between TFs and chromatin modifiers.\(^\text{99,100}\) The impact of TF activities on epigenome organization is probably best illustrated by cell-state reprogramming. An adult cell can be reprogrammed to an induced pluripotent stem (iPS) cell by ectopic expression of four TFs: Oct4, Nanog, Klf4, and Myc.\(^\text{101}\) The reprogramming
process is accompanied by dramatic epigenetic changes. Most significantly, the DNA methylation marks at the pluripotent regulators are erased and bivalent domains are reestablished. All four reprogramming factors interact with chromatin modifiers. Specifically, Oct4 interacts with the SWI/SNF complex member Wdr5;69 Nanog interacts with the histone deacetylase NuRD and SWI/SNF complex,102 Klf4 can interact with the histone acetyltransferase p300;103 and most strikingly, Myc, can interact with many chromatin factors such as histone acetylases (GCN5, p300), chromatin-remodeling complexes, histone deacetylases, and histone demethylases. It remains unclear to what extent the global epigenetic changes are directly caused by these TF-chromatin interactions.

Computational studies of various epigenomic datasets have suggested that, in general, the contribution of each individual TF in maintenance of the epigenetic patterns is typically not significant. For example, for nucleosome positioning, there are at least three TF motifs that are enriched in nucleosomal sequences: ABF1, REB1, and STB2.24 For DNA methylation, the preference seems to be dictated by interaction with histone modifying complexes and epigenetic patterns is mainly contributed by a few features that are traditionally viewed as degenerative, whereas the sophisticated models described above often lead to quantitative but not qualitative improvement. For example, while many methods have been developed to predict nucleosome positions, it seems that the overall accuracy is only moderately higher than a simple model using the G+C density as the single predictor.62 Similarly, for DNA methylation, the genome-wide pattern is most strongly correlated with the CpG density, whereas including other features such as TF motifs and repetitive sequences, only adds moderate prediction power.19–21 As discussed above, the CpG density is also strongly associated with histone modifications such as H3K4me3 and H3K27me3.

Interestingly, eukaryotic cells have developed various mechanisms to recognize these general features. For nucleosome positioning, the G+C content can directly affect the nucleosome-DNA binding affinity. For different histone modifications, the sequence signals seem to be mediated by different factors. The recruitment of H3K4me3 is mediated by the CXXC domain-containing protein Cfp1,68 which recognizes unmethylated CpG. The association between PcG complexes and CpG islands may be mediated by interaction with Jarid2,104–107 which binds DNA with bias toward G+C rich sequences.104 It has been proposed that the PcG complexes target CpG islands by default, and this default recruitment is antagonized by the simultaneous recruitment of activators by specific TF motifs.60 Furthermore, cross-talk between these machineries is likely to play an important role in establishing the overall epigenetic pattern. For example, for DNA methylation, the preference seems to be dictated by interaction with histone modifying complexes such as G9a and PcG proteins.14,108

Noncoding RNA

NcRNAs are an emerging class of potentially important regulators.109,110 Two types of ncRNAs, long noncoding RNAs (lncRNA) and short RNAs, have been implicated in chromatin recruitment. Many short RNAs are involved in the RNAi pathway, which plays an important role in heterochromatin formation. An unrelated class of short RNAs have been found to be transcribed from the 5′ end of PcG target genes,111 and these short RNAs form a stem-loop structure that is able to physically interact with PRC2. In addition, thousands of lincRNAs have been detected in mammalian cells.112 About 20% of these transcripts can physically interact with PRC2.113 While the function of most lincRNAs remains uncharacterized, a few examples, such as Xist and HOTAIR, have been investigated intensively. Xist is the major regulator for X-inactivation and able to recruit PcG complexes to the parent locus, to which it remains tethered.79 The 5′ end of Xist contains a RepA domain which can interact with PRC2.79 The ability to interact with both PRC2 and DNA makes Xist act as a linker between them. HOTAIR also physically associates with PRC2,78 but its target region is located on a different chromosome, and the underlying mechanism remains unclear.

A major challenge to further investigate the role of ncRNAs is to understand their targeting mechanism. As discussed above, while some ncRNAs act in cis, others target distant regions that may be located on a different chromosome. Several intriguing models for trans-action have been recently proposed,113 including (1) forming a RNA–DNA triplex; (2) serving as a link by interaction with both chromatin complexes and TFs; (3) inducing specificity of a chromatin remodeling protein through allosteric modification; and (4) mediating long-range chromatin interactions.

© 2012 Wiley Periodicals, Inc.
DNA SEQUENCE ASSOCIATED WITH OVERALL EPIGENETIC VARIABILITY

While epigenetic patterns are cell type specific, the variability is not uniform across the genome. A growing body of literature suggests that the genome-wide distribution of epigenetic variability is closely associated with the underlying DNA sequence. In the following I briefly review some recent results.

Nucleosome Positioning

In yeast, most promoters are associated with NFRs independent of the transcription status. These promoters are typically TATA-less. In contrast, the nucleosome positions in the TATA-box containing promoters are dynamic depending on the growth conditions. Under an inducible condition, nucleosomes can be evicted by SWI/SNF, thereby making the DNA regulatory elements accessible for TF binding followed by gene activation. In human and mouse, it has been shown that eviction of nucleosomes containing the H3K4me2 marks are generally associated with active enhancers. This relationship is likely to be valid in a wide range of species rather than mammal specific. Interestingly, the insulator protein CTCF targets DNA sequences with intrinsically high nucleosome occupancy. Once bound, CTCF displaces neighboring nucleosomes, thereby creating a regular nucleosome array.

Experimental techniques have also been developed to measure the histone turnover rate. These investigators labeled ancestral and newly synthesized histones with different tags and then measured the enrichment ratio of two tags on a genome-scale. They found that, in cells that are arrested in the G1 stage of the cell-cycle, the histone turnover rate is higher in promoter regions and chromatin boundary elements. In dividing cells, the ancestral histones are redistributed as a result of DNA replication. Surprisingly, the ancestral histones are not reincorporated at the same location but can be displaced within a window of ~400 bp which results in accumulation of ancestral histones at 5′ end of genes.

Histone Modification

Changes of the histone modification pattern in promoter regions are directly related to gene regulation. For both H3K4me3 and H3K27me3, the degree of cell type specific variability is highly associated with its CpG density. In particular, High CpG promoters (HCP) are highly associated with H3K4me3, regardless the transcription activities. Dynamic changes are mostly associated with low CpG promoters (LCP). The variability of H3K27me3 is also associated with CpG content. During cell differentiation, the loss of focal peaks is compensated by expansion of large-scale domains. As a result, the overall occupancy is nearly preserved. In addition to the CpG density, a handful of TF motifs (such as E2F, ZF5, SP1, and MYC) can also contribute to the overall variability.

The overall variability of promoter histone modification patterns is surprisingly low. Many genes that are expressed only in certain cell types are already associated with open chromatin at the promoter regions. On the other hand, the patterns are much more dynamic at enhancers, and such variation is highly correlated with gene activities. Enhancers are often enriched with highly conserved noncoding elements and contain clusters of TF motif sites.

DNA Methylation

The variability of DNA methylation is also highly associated with the CpG density. During normal cell differentiation, dynamic changes are most concentrated on LCPs, while most HCPs are unmethylated across different cell types. Indeed, the hypermethylated HCPs also tend to have lower CpG content than the unmethylated ones. Recent studies have shown that this relationship also holds on the genome-scale. Large-scale aberrant methylation patterns have been long recognized in cancer and may play an important step toward DNA tumorigenesis. A recent study has identified significant overlap between regions that are differentially methylated among tumors and those among different normal tissues, suggesting a developmental origin of these aberrations. Interestingly, these studies have also observed large-scale domains of high variability. While it is unclear how these domains are established, it is possible that certain sequence features, such as the CTCF motif sites, may play an important role in maintenance of the boundaries.

DISCUSSION

What has emerged from these studies is a consensus that the overall epigenetic pattern is not independent of, but closely associated with the genomic sequence (Figure 1). The DNA sequence seems to play at least three important roles: (1) Generic features, such as the G+C content and CpG density, may be used to shape the skeleton of the epigenomic patterns, by facilitating various protein-DNA interactions mostly involving a small set of general factors; (2) Specific features, such as TF motifs, can recruit TFs, ncRNAs, and other factors in a cell type specific manner, but the immediate impact of each factor may be limited...
FIGURE 1 | A model for the role of the genomic sequence in guiding genome-wide epigenetic patterns. (a) The intrinsic association between an epigenetic mark and a genome locus is sequence-dependent and can be quantified by a propensity score. (b) An epigenetic mark (represented by the stars) is constitutively recruited to regions with high propensity scores, mediated by interaction with general factors (represented by the ovals) that target distinct sequence features. These features are usually degenerative and are the main determinants of the propensity scores. On the opposite end, this mark is excluded from regions with low propensity score. In the middle range, the epigenetic pattern is highly variable among different cell types. Occupancy can be enhanced or inhibited because of interactions with many cell type specific factors (represented by other shapes).

to a specific biological pathway or cell type; and (3) the DNA sequence information may also be used to orchestrate the overall epigenetic variability, such that certain regions are more plastic than others. One likely scenario is that, while the general factors described above may play a major role in initializing the epigenetic patterns at an early developmental stage, cell type specific factors, such as TFs and ncRNAs, may play a more important role later on by mediating the divergence of epigenomic patterns across different cell types.

While I have mainly focused on the DNA sequence, it is important to recognize that this is only one component of the complex network of diverse interactions that regulates global epigenetic patterns. For example, once initialized, the epigenetic marks may serve as a platform that can be used to recruit additional marks, thereby serving as a self-enhancing feedback loop, which may also lead to significant spatial expansion, resulting large-scale domains spanning hundreds of kilobases.6,9,67,130 On the other hand, different epigenetic marks compete through cross-talk between chromatin readers, effectors, and erasers, and between these machineries and DNA sequence binding factors.131 Long-range chromatin interactions may play an important role in maintenance of domain boundaries.132 All in all, the genome-wide epigenetic pattern should not be viewed as a simple equilibrium but a steady state resulting from the delicate balance of these complex interactions. An important gap in our current knowledge is how distinct combinatorial epigenetic patterns are established, which is not clear even for promoter regions.133–135 We are just beginning to understand this complex network of these interactions (see Refs 13–15 for more comprehensive reviews). Future investigations will benefit from integration of diverse data-types.9

Systems biology studies have shown the expression level of many genes are not precisely controlled but undergo significant stochastic fluctuation.136 It has been theorized that phenotypic diversity is evolutionarily beneficial for robustness against environmental perturbations.129,137–139 It has been increasingly recognized that epigenetic regulation plays an important role in the maintenance of gene expression diversity. It is very interesting that the degree of epigenetic variability is associated with certain DNA sequence signatures, providing a simple mechanism for inheritance. As suggested by several recent studies,139,140 elucidating the connection between DNA sequence in epigenetic variability will provide mechanistic insights into the evolutionary history of phenotypic diversity.

ACKNOWLEDGMENTS

I apologize for omission of relevant references due to space restrictions. I thank Drs Franziska Michor and Shirley Liu for helpful discussions. This research was supported by an HSPH Career Incubator Award and the NIH grant 1R01HG005085-01A2.

REFERENCES


