Chapter 8

Prediction of Epigenetic Target Sites by Using Genomic DNA Sequence

Guo-Cheng Yuan
Harvard School of Public Health, USA & Dana-Farber Cancer Institute, USA

ABSTRACT

Epigenetic regulation provides an extra layer of gene control in addition to the genomic sequence and is critical for the maintenance of cell-type specific gene expression programs. Significant changes of epigenetic patterns have been linked to developmental stages, environmental exposure, ageing, and diet. However, the regulatory mechanisms for epigenetic recruitment, maintenance, and switch are still poorly understood. Computational biology provides tools to deeply uncover hidden connections and these tools have played a major role in shaping the current understanding of gene regulation, but its application in epigenetics is still in the infancy. This chapter reviews some recent developments of computational approaches to predict epigenetic target sites.

INTRODUCTION

Epigenetics refers to heritable changes of gene expression or genotypes without change of the DNA sequence (Waddington, 1942). In a multicellular organism, the DNA sequence is constant in all cell lineages, but the gene activities in different cell-types are highly variable. Such cell-type specific gene expression patterns are controlled by epigenetic mechanisms, including nucleosome positioning, histone modifications, and DNA methylation. Together these mechanisms control the accessibility of the genomic DNA to regulatory proteins. Only a small part of the genomic blueprint is used in any cell type.

DOI: 10.4018/978-1-60960-491-2.ch008

The rapid advance of microarray and DNA sequencing technologies (Barski et al., 2007; Mikkelsen et al., 2007; Ren et al., 2000) has allowed researchers to identify genome-wide epigenetic patterns in various species. Recent epigenomic studies have identified dramatic epigenetic differences between different cell-types (Barski et al., 2007; Heintzman et al., 2009; Meissner et al., 2008; Mikkelsen et al., 2007; Mohn et al., 2008), between normal and disease tissues (Schlesinger et al., 2007; Seligson et al., 2005; TCGA, 2008), and between stimulated and resting cells (Saccani & Natoli, 2002; Wei et al., 2009). These differences are also highly correlated with gene expression level changes. Importantly, the epigenetic changes are not permanent but can be reversed. Strikingly, the entire epigenetic state in an adult cell can be reprogrammed to a pluripotent cell state (called iPS cell) that is highly similar to an embryonic stem (ES) cell (Okita et al., 2007; Takahashi & Yamanaka, 2006; Wernig et al., 2007; Yu et al., 2007). This reversibility makes epigenetic marks the ideal targets for therapeutic treatment (Sharma et al., 2010). Indeed, a number of drugs have been developed and currently used to treat a number of diseases including cancer (Yoo & Jones, 2006). However, a major challenge is to avoid off-target interactions, since our current understanding of the targeting mechanisms of epigenetic factors is still limited.

The epigenetic pattern is not randomly distributed across the genome (Bernstein et al., 2007). A fundamental question is how target specificity is achieved. The targeting mechanism is complex and involves many factors such as the genomic DNA sequence, chromatin modifiers, transcription factors (TFs), and non-coding RNAs. How these factors work together to regulate epigenetic patterns is still poorly understood. Among these factors, the association with DNA sequence has been most studied. Perhaps the most commonly studied factor is the DNA sequence. Here I review some of computational studies aimed at prediction of epigenetic patterns based on the DNA sequence. Additional information on certain specific aspects can be found in some excellent reviews (Bock & Lengauer, 2008; Kaplan et al. 2010). In addition, the readers are referred to some excellent reviews for biological background (Jiang & Pugh, 2009; Kouzarides, 2007; Rando & Chang, 2009; Hawkins et al. 2010; Zhou et al. 2011).

METHODS TO PREDICT EPIGENETIC TARGETS

Nucleosome Positioning

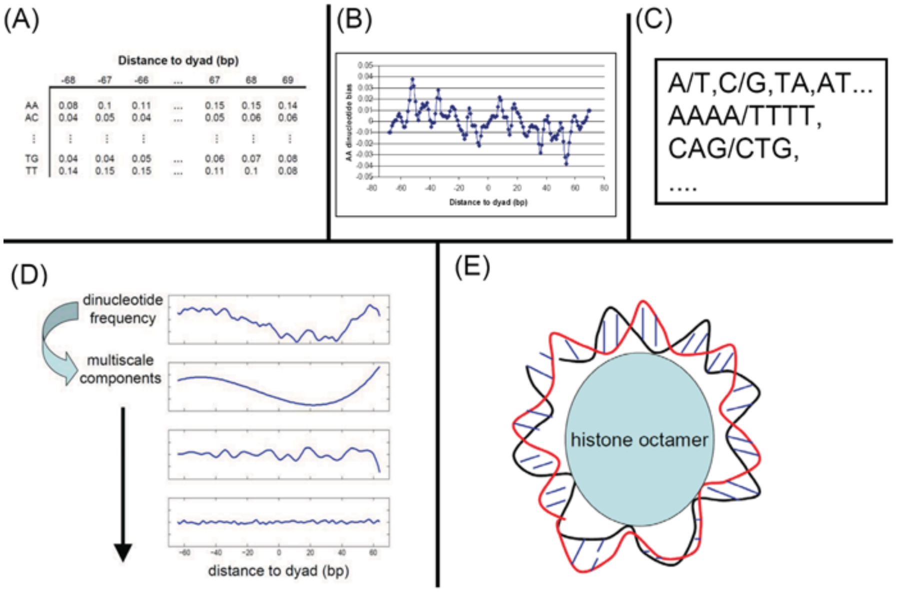
The eukaryotic DNA is packaged into chromatin. The fundamental unit of chromatin is the nucleosome, consisting of two copies each of four core histone proteins: H2A, H2B, H3, and H4 (Kornberg & Lorch, 1999). Each nucleosome wraps around 146 bp DNA in about 2 turns. Ever since the initial discovery of nucleosomes in the 70's (Kornberg, 1974), the regulatory mechanisms underlying nucleosome positioning have been intensely investigated. A potentially important role of DNA sequences was noticed decades ago. The investigators recognized that the structural properties of DNA are dependent upon the base pair composition; therefore specific DNA sequences might be favored for nucleosome binding. By extracting nucleosome core particles from chicken red blood cells and then analyzing the DNA sequences attached to these nucleosome particles, Satchwell et al. (1986) observed an approximately 10 bp periodic pattern of the frequency of the dinucleotide pair AA/TT. Similar results were found by several other groups (Ioshikhes et al., 1996; Widom, 2001). The 10 bp periodicity pattern agrees well with the high-resolution nucleosome structure (Luger et al., 1997; Richmond & Davey, 2003), where the histones interact with the DNA sequence approximately once every 10 base pair.

However, early studies were limited by the fact that only a handful of sequences were known to be nucleosome bound. During the past decade, rapid progress has been made thanks to the development of high throughput technologies such as microarrays and DNA sequencing. As a result, high-resolution, genome-scale maps of nucleosome positions have been identified in various organisms including human (Chodavarapu et al. 2010; Johnson et al., 2006; Lanterman et al. 2010; Lee et al., 2007; Mavrich, Ioshikhes et al., 2008; Mavrich, Jiang et al., 2008; Ozsolak et al., 2007; Schones et al., 2008; Yuan et al., 2005). Despite the wide-range of species, the overall nucleosome positioning pattern is strikingly conserved. A common signature at the promoter regions is characterized by a nucleosome-free region (NFR) adjacent to the transcription start sites (TSS) flanked by well-positioned nucleosomes.

Segal et al. (2006) were the first to develop computational methods to predict genome-wide nucleosome positions (Figure 1A). Based on a set of 199 high-resolution nucleosome DNA sequences obtained by direct sequencing, the authors constructed a position specific weight matrix to model the nucleosome-DNA binding affinity. The approach is similar to standard methods for TF motif detection, but an important difference is that the basic units are dinucleotides (AA, CpG, etc.) instead of single nucleotides (A, C, G, and T). Consistent with previous studies, the average AA/TT/TA pattern is approximately periodic with about 10 bp periodicity. In order to predict genome-wide nucleosome positions, Segal et al. (2006) further developed a thermodynamic model, which took into account the steric hindrance effect between neighboring nucleosomes. Impressively, their model is able to predict 54% of nucleosome positions within 35 bp accuracy. However, the significance of this result is compromised by the observation that even random guesses could provide 38% prediction accuracy when evaluated in the same way. The underlying problem with this accuracy measure is that it increases with the total number of predicted sites, making it difficult to interpret the results. Around the same time, Ioshikhes et al. (2006) used a different approach to predict genome-wide nucleosome positions. They computed the correlation between an input sequence and a pre-determined nucleosome positioning sequence (NPS) pattern and then detected regions of high correlation score (Figure 1B). They observed that the two approaches shared similar prediction accuracy.

The above studies considered only information from the nucleosome sequences. On the other hand, it has been recognized that certain DNA sequences such as poly dA:dT runs are inhibitive for nucleosome binding (Bernstein et al., 2004; Sekinger et al., 2005; Yuan et al., 2005). Such sequences impose an important constraint on the nucleosome positions. Recognizing the role of nucleosome-free (linker) sequences, several groups have developed computational models to incorporate both nucleosome and linker DNA sequences for prediction (Lee et al., 2007; Peckham et al., 2007; Yuan & Liu, 2008), but the classification strategies are quite different. Specifically, Peckman et al. (2007) represented a sequence pattern by the number of word counts corresponding to a set of short *k*-mers (*k* up to 6), and the differences between nucleosome and linker sequences were detected by using a support vector machine (SVM) (Figure 1C). Lee et al. (2007) characterized sequence signatures based on TF motif scores and DNA structural parameters and used a Lasso regression method as a classifier. A third approach converted the dinucleotide frequencies to wavelet coefficients in order to detect discriminative periodic patterns, and a stepwise logistic regression model was used to distinguish nucleosome bound sequences from those located in the NFRs (Yuan & Liu, 2008) (Figure 1D). In the latter study, it was also found that a more objective measure for model performance was the false positive error rate rather than the false negative error rate which were used in previous studies. Despite the variation of the predictor selection and classification schemes, these models all significantly improve the model performance. More recently, Segal and colleagues

Figure 1. Schematic diagrams to illustrate the concepts behind various nucleosome positioning prediction methods. (A). Segal et al. (2006) modeled the nucleosome sequence pattern by using a position specific weight matrix. Each entry represents the probability of observing a specific dinucleotide at a specific position. (B) Ioshikhes et al. (2006) defined a NPS pattern based on the bias of AA/TT distribution. (C) Peckham et al. (2007) extracted a large number of sequence features by counting the occurrence of various short words. (D) Yuan and Liu (2008) studied the role of periodicity by decomposing a multiscale signal to various wavelet components, each varying at a specific length scale. (E) Miele et al. (2008) and Morozov et al. (2010) calculated the free energy required for any DNA sequence to wrap around a nucleosome. A sequence associated with lower energy is more favored for nucleosome binding.



have also incorporated negative controls in their model framework and obtained much improved performance (Field et al. 2008).

The methods mentioned above are all based on empirical data. While these models can provide good model accuracy, they do not necessarily offer mechanistic insights. In the meantime, a different class of models has been recently developed based on calculation of the biophysical properties (Miele et al., 2008; Morozov et al., 2009) (Figure 1E). Both studies showed that a biophysically-based model may offer competitive performance for genome-wide predictions.

It is also important to note that, just because the DNA sequence and nucleosome positions are correlated, it does not mean that it is deterministic. In fact, Kornberg and Stryer (1988) pointed out that positioned nucleosomes can also be predicted based on a statistical model. In this model, only the nucleosome boundaries are determined by the DNA sequences, whereas the nucleosomes themselves are randomly packed. This mechanism results in highly aligned nucleosomes near the boundary whereas fuzzier configuration elsewhere, which is in fact consistent with experimental data (Mavrich, Ioshikhes et al., 2008). In addition, the *in vivo* nucleosome positions are inevitably affected by additional factors, such as the perturbation by chromatin remodeling complexes, competition with TF binding, and influence of transcriptional events, it is unclear to what extent the resulting positions are intrinsically coded in DNA sequences. To overcome this challenge, two groups (Kaplan et al., 2009; Zhang et al., 2009) have recently used next generation DNA sequencing methods to map the *in vitro* nucleosome positioning for sequences extracted from the yeast genome. Interestingly, these studies draw different conclusions although their data are similar. Kaplan et al. (2009) concluded that the intrinsic DNA sequence preference of nucleosomes have a central role in determining nucleosome positioning *in vivo*, noting that the nucleosome occupancy level is similar between *in vivo* and *in vitro* environments. On the other hand, Zhang et al. (2009) concluded against a genomic code for nucleosome positioning, pointing out the observed similarity can be simply explained by a statistical positioning model (Kornberg & Stryer, 1988).

Histone Modification

The N-terminal ends of the core histone proteins are unstructured and referred to as the histone tails, which can be post-translationally modified in multiple ways at multiple sites. Early studies were focused on histone acetylation, the modification that an acetyl group is added to a lysine residue. While promoter histone acetylation is generally associated with gene activation (Dion et al., 2005; Roh et al., 2005), histone acetylation in coding regions may lead to gene repression (Wang et al., 2002). Another well-characterized mark is histone methylation. The function of histone methylation is more complex and still not well understood. For example, H3K9 methylation is highly correlated with gene repression but H3K4 methylation is correlated with gene activation (Barski et al., 2007; Pokholok et al., 2005). An additional complexity for histone methylation is that it can happen in three flavors: mono-, di-, and tri-methylation, each may have its own role. For example, H3K4me3 is highly correlated with active promoters (Barski et al., 2007; Guenther et al., 2007), whereas H3K4me1 tends to be depleted at promoters but enriched at tissue-specific enhancers (Heintzman et al., 2007). In addition to acetylation and methylation, a large number of histone modifications have been identified including phosphorylation, ubiquitylation, ADP ribsylation, deimination, and praline isomerization (Kouzarides, 2007). The functionality for most of these modifications is still poorly understood and their function may be also context dependent. The task for identifying the function of the combinations of different histone modification marks is generally referred to as the "histone code" hypothesis, originally proposed by Allis and colleagues (Jenuwein & Allis, 2001; Strahl & Allis, 2000).

Although histone modifying enzymes do not interact with the DNA sequence, they may be recruited to specific loci by interacting with TFs, non-coding RNAs, or other DNA interacting regulators. There is abundant evidence that the distribution of many histone modification marks is associated with the DNA sequence each in its own way (Bernstein et al., 2007). For example, the H3K4me3 mark is mainly associated with high density and can be well-predicted by a CpG density alone (Bernstein et al., 2006), while the H3K9me3 mark is weakly associated with a number of repetitive sequences (Martens et al., 2005).

Computational studies for histone modifications have been mainly limited to H3K27me3 through investigation of Polycomb group (PcG) proteins targeting. PcG was first discovered in *Drosophila* for controlling Homeotic (Hox) genes (Lewis, 1978) but later found to also play an important role in early development in vertebrates (Schuettengruber et al., 2007; Sparmann & van Lohuizen, 2006). A major function of PcG proteins is to repress the transcriptional activities of their target genes through tri-methylation of the histone H3 on lysine 27 residue (H3K27me3) (Francis & Kingston, 2001). Much effort has been taken to identify the DNA elements that are responsible for PcG recruitment, called the Polycomb response elements (PRE). In *Drosophila*, Ringrose et al. showed that the PREs are well-characterized by the motif sequences of distinct TFs among which

PHO is the most important one (Ringrose et al., 2003). On the other hand, the mammalian PREs are still poorly characterized (Schuettengruber et al., 2007; Simon & Kingston, 2009). Experimentally validated PREs have only been recently identified (Sing et al., 2009; Woo et al., 2010).

Computational predictions of PREs are mainly centered around TF motifs. For *Drosophila*, Ringrose et al. found that individual TF motifs are insufficient for discriminating PREs from non-PREs (Ringrose et al., 2003). However, by pairing different motifs together, the discriminative power was much improved. These authors then derived a score based on a linear combination of the motif-pair scores. To test their prediction accuracy, they experimentally validated 43 regions randomly selected from a total of 167 predicted sites. 29 out of the tested regions were verified. More recently, additional TF motifs were incorporated in the same modeling framework, which led to improved model performance (Hauenschild et al., 2008).

In comparison, mammalian PREs are less characterized. Ku et al. (2008) investigated the association between TF motifs and genome-wide PcG targets in mouse ES cells. They identified several distinct TF motif patterns and used these patterns to predict PcG targets. Among the top 2836 predicted targets, about 60% are correct predictions. Similar results were obtained by using a more sophisticated model (Liu et al., 2010). In this study, Liu et al. (2010) found that the highly scoring genes tend to be marked by PcG in multiple cell-types, suggesting the DNA sequence is strongly related to target plasticity.

Currently, general methods for histone modification target predictions are still limited. A major challenge is that there are a large number of possible combinations, each has its own distribution profile. Some are focal (e.g. H3K4me3), others are broader (e.g. H3K9me3), and yet others are mixed (e.g. H3K27me3) (Barski et al., 2007). Another complexity is that the factors that regulate histone modifications are more complex. A histone modification mark can be either added or removed by specific enzymes. There are a large number of such enzymes, many of which share overlapping roles (Kurdistani & Grunstein, 2003; Lan et al., 2008). Since each factor functions differently, it is likely each only contributes to a small subset of targets.

A modified version of the wavelet model mentioned above has been applied to predict histone modification patterns in human (Yuan, 2009). The model performance is highly variable among different histone modification marks. For a few well-studied histone modification marks, such as H3K4me3 and H3K4me1, the model indeed performs well and the performance cannot simply be explained the local enrichment of CpG. On the other hand, the model predicts H3K9me3 rather poorly. The performance of the model is correlated with the overall spread of a histone modification mark. Interestingly, the H3K4me2 and H3K27me3 marks do not overlap in adult cells, yet their target sequences are highly similar. A possible explanation is that the two sets of marks both target same regions but only one mark can be established. Experimental evidence supporting this possibility is that the H3K4me2 pattern at the HoxA cluster in one tissue (lung) is similar to the H3K27me3 pattern in a different tissue (foot) (Rinn et al., 2007), suggesting that the targeted competition between the two marks may be responsible for epigenetic switching.

DNA Methylation

The genomic DNA itself can also be covalently modified and the modification has important implication on gene regulation (Bird, 2002). In this case, the cytosine nucleotide can be methylated. With the exception of a few special cases (Cokus et al., 2008; Lister et al., 2009), DNA methylation almost always occurs in the context of a CpG dinucleotide. Since CpG dinucleotide is self-complementary, the DNA methylation pattern on one DNA strand can be faithfully reproduced on the other strand, a property that is important for epigenetic inheritance. While promoter DNA methylation is often correlated with gene repression, recent epigenomic studies have shown that DNA methylation can also occur at coding region and its functional consequence is still poorly understood. In cancer, it has been found that the genome-wide DNA is widely hypomethylated, whereas specific loci, such as certain tumor repressor genes, are associated with hypermethylation (Esteller, 2007; Jones & Baylin, 2007). The overall methylation level can be influenced by food intake such as folic acid (Jirtle & Skinner, 2007).

Not surprisingly, the DNA methylation status is highly correlated with the local CpG density. The majority of CpG is located in low CpG density regions and tends to be methylated. On the other hand, CpG can also form clusters called the CpG islands, which tend to be unmethylated. However, some CpG islands are methylated in certain cell-types but not others. The sequence characteristics of such differentially methylated regions (DMR) are still incompletely understood, although it has been shown that DMRs are typically associated with intermediate CpG density (Bock et al., 2006). In cancers, it was found that a number of tumor repressor genes are silenced by DNA methylation (Keshet et al., 2006; TCGA, 2008). A challenge is to understand which set of CpG islands can be methylated.

A number of computational methods have been developed to predict the CpG island methylation from the underlying DNA sequence (Bock et al., 2006; Das et al., 2006; Fang et al., 2006; Feltus et al., 2006; Keshet et al., 2006). The overall strategies in these studies are similar, although there is a variation of the predicting sequence features that are used in these studies. For example, Das et al. (2006) used 102 sequence features as predictors, including GC content, word counts, and repetitive sequences. The classification was done by using support vector machine. Feltus et al. (2006) obtained discriminative sequence pattern by de novo motif searching followed by a decision tree as the classification model. In addition, Bock et al. (2006) incorporated the DNA structural parameters as predictors. In this study, the authors also used a similar approach to predict several other epigenetic marks including histone modifications and DNA hypersensitivity and then combine the information together to predict the overall strength of a CpG island. The biological interpretation of the strength of a CpG island is the likelihood of being kept in an open chromatin state and targeted by TFs. Recent analysis has shown that integrating the histone modification pattern information can improve model performance (Fan et al.).

DISCUSSION

The Role of DNA Sequence in Defining Epigenetic Patterns

The targeting mechanism for epigenetic factors is complex and involves a large number of factors and this complexity is only beginning to be investigated systematically. Here we discuss an important first step, which is the role of DNA sequence in shaping the global epigenetic landscape. The results reviewed in this paper strongly indicate that the DNA sequence plays an important role in the targeting of many epigenetic marks. There are some important exceptions. For example, the H3K36me3 pattern is mainly determined by transcription rather than coded in the DNA sequence.

Although the detailed mechanism is still unclear, we can think of the DNA sequence as defining the intrinsic stability of an epigenetic mark. In the case of nucleosome positioning, the predicted stability has been directly validated by genetic experiments and it is found that the nucleosome occupancy indeed change as predicted (Segal et al., 2006; Sekinger et al., 2005). These results suggest that the DNA sequences are indeed required for establishment of the proper nucleosome positions. Interestingly, the DNA sequence at the 5' end of a coding region is typically coded for high nucleosome occupancy, which may act as important barrier for passage of transcriptional machineries. Recent studies have found that PolII occupies many inactive genes but only is paused near TSS (Core & Lis, 2008; Guenther et al., 2007), whereas can be paused at the 5'end and used only to generate short incomplete transcripts (Core & Lis, 2008; Guenther et al., 2007). The requirement to overcome this barrier to finish a full transcript suggests an important transcription control mechanism (Mavrich, Ioshikhes et al., 2008).

Although less understood, the DNA sequence is also related to the overall plasticity of other epigenetic marks such as DNA methylation and histone modification. For DNA methylation, the regions with high CpG density tend be unmethylated, whereas those associated with low CpG tend to be methylated. Interestingly, the most variable regions seem to be related to intermediate CpG content (Bock et al., 2008; Das et al., 2006; Feinberg & Irizarry, 2010). A similar but more intricate pattern has been found for histone medications as well.

Recent studies have found that cancer is not only characterized with high genetic changes but also high epigenetic changes (TCGA, 2008). Interestingly, the aberrant DNA methylation pattern is correlated with genetic mutations. For example, in treated samples which display DNA methylation at the MGMT promoter, 81% of all mutations are of the G:C to A:T type in non-CpG dinucleotides, compared to a mere 4% within CpG dinucleotide. In comparison, in samples without MGMT methylation, the frequencies of the two types of mutations are roughly equal (29% vs 23%, respectively). It is still unclear whether other epigenetic changes are also correlated with genetic mutations in cancer.

Finally, we recognize that a lot of work is still needed to gain mechanistic insights. For example, despite the success of DNA sequence in prediction of nucleosome occupancy, the *in vivo* nucleosome positioning pattern can be simply explained by a statistical positioning model (Kornberg & Stryer, 1988; Mavrich, Ioshikhes et al., 2008; Zhang et al., 2009), suggesting that the DNA sequence may only be important for delineating the boundaries of nucleosome occupied regions.

Beyond the Sequence

While the DNA sequence is constant across cell-types, the actual epigenetic pattern is tissue-specific and cannot be determined by the DNA sequence alone. There are a large number of potential regulators, including chromatin modifying enzymes, TFs, and non-coding RNAs. For example, the ATP dependent chromatin remodelers can remove nucleosomes from their favored positions. A classical example is the regulation of PHO5 (Svaren & Horz, 1997). At normal conditions, the PHO5 promoter is occupied by well-positioned nucleosomes, one of which is centered at -275 bp relative to the ATG codon, occluding Pho4 from binding to its target site at -247 bp (Almer et al., 1986). This and three other nucleosomes are depleted upon phosphate starvation, making the Pho4 binding site accessible. The eviction of nucleosome is caused by the activity of SWI/SNF, an ATP-dependent chromatin remodeler. Similarly, the tissue-specific patterns of histone modification and DNA methylation are also highly dependent on the activities of various histone and DNA modification enzymes.

Since chromatin modifiers can be recruited to specific target sites by interacting with sequence-specific TFs. The activity of these TFs can also significantly affect the overall epigenetic pattern. For example, the histone deacetylase Hst1 is recruited by a single TF Sum1 in yeast (Robert et al., 2004). The genome-wide targets of Hst1 and Sum1 are nearly identical. Genetic deletion of SUM1 completely abolishes the binding of Hst1 and causes increased H3 and H4 acetylation level at their target sites. Similarly, in *Drosophila*, the TF PHO plays an important role in PcG recruitment, and deletion of PHO results in derepression of Hox genes, indicating disrupted PcG binding (Wang et al., 2004).

Another class of regulators that have been recently described is the non-coding RNAs (Guttman et al., 2009; Rinn et al., 2007; Zhao et al., 2008). For example, in mammals one of the X chromosomes in females is completely silenced for dosage compensation. This whole chromosome silencing is mediated by the DNA methylation. The large Xist RNA is produced at the inactive X-chromosome and thought to initiate the establishment of DNA methylation and X-inactivation (Lee, 2009). In addition, small RNAs can also interact with chromatin modifiers thereby regulating the local histone modification and DNA methylation patterns (Moazed, 2009).

These examples demonstrate that there are a large number of potential regulators for epigenetic patterns. A fundamental task is to understand their respective roles in establishing the global epigenetic patterns. Computational methods suitable for this task have yet been developed.

Epigenetics and Evolution

Modern evolutionary theory is firmly based on genetic variation and natural selection. The role of epigenetics in evolution remains unclear. Feinberg and Irizarry have recently proposed that the stochastic epigenetic variation originated from genetic variation may play an important role in evolutionary adaptation (Feinberg & Irizarry, 2010). Such variation does not change the mean phenotype, but stochastic variation is advantageous for adaptation to environmental changes. By using numerical simulation, the authors demonstrated that the increased variation can indeed increase fitness in a varying environment. They also found experimental evidence supporting that the locations of variably methylated regions across different samples are correlated with the local CpG density. Interestingly, the variability changes between human and mouse accompanied by CpG density changes. A core component of this hypothesis is that genetic variation is closely related to epigenetic variation, which is supported by the numerous studies reviewed in this paper.

Several studies have taken a comparative genomic approach to investigate whether genomic variations may be associated with expression changes via difference in epigenetic patterns. TF binding sites are found to be typically associated rigid DNA (Tirosh et al., 2007), consistent with nucleosome depletion at these sites. Interestingly, these authors also found that the locations of rigid DNA elements are conserved in TATA-less promoters but vary substantially at the promoters containing the TATA element. These differences are thought to be related to be developed during evolution to initiate species-specific responses to environmental changes. Along the same line, Field et al. predicted the promoter nucleosome occupancy level for various yeast species based on the genomic sequences (Field et al., 2009). Interestingly, the predicted nucleosome occupancy level is substantially different at genes which have different expression patterns between different yeast species. In particular, the respiratory genes are active in aerobic yeast species but inactive in anaerobic ones. In accordance to this difference, the predicted nucleosome level is low at the respiratory promoters for the aerobic species but much higher in other ones.

ACKNOWLEDGMENT

This research was supported by a Claudia Adams Barr Award.

REFERENCES

Almer, A., Rudolph, H., Hinnen, A., & Horz, W. (1986). Removal of positioned nucleosomes from the yeast PHO5 promoter upon PHO5 induction releases additional upstream activating DNA elements. *The EMBO Journal*, *5*(10), 2689–2696.

Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., & Wang, Z. (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, *129*(4), 823–837.

Bernstein, B. E., Liu, C. L., Humphrey, E. L., Perlstein, E. O., & Schreiber, S. L. (2004). Global nucleosome occupancy in yeast. *Genome Biology*, *5*(9), R62.

Bernstein, B. E., Meissner, A., & Lander, E. S. (2007). The mammalian epigenome. *Cell*, *128*(4), 669–681.

Bernstein, B. E., Mikkelsen, T. S., Xie, X., Kamal, M., Huebert, D. J., & Cuff, J. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, *125*(2), 315–326.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes & Development*, *16*(1), 6–21.

Bock, C., & Lengauer, T. (2008). Computational epigenetics. *Bioinformatics (Oxford, England)*, *24*(1), 1–10.

Bock, C., Paulsen, M., Tierling, S., Mikeska, T., Lengauer, T., & Walter, J. (2006). CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. *PLOS Genetics*, *2*(3), e26.

Bock, C., Walter, J., Paulsen, M., & Lengauer, T. (2008). Inter-individual variation of DNA methylation and its implications for large-scale epigenome mapping. *Nucleic Acids Research*, *36*(10), e55.

Chodavarapu, R. K., Feng, S., Bernatavichute, Y. V., Chen, P. Y., Stroud, H., & Yu, Y. (2010). Relationship between nucleosome positioning and DNA methylation. *Nature*, *466*(7304), 388–392.

Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., & Haudenschild, C. D. (2008). Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, *452*(7184), 215–219.

Core, L. J., & Lis, J. T. (2008). Transcription regulation through promoter-proximal pausing of RNA polymerase II. *Science*, *319*(5871), 1791–1792.

Das, R., Dimitrova, N., Xuan, Z., Rollins, R. A., Haghighi, F., & Edwards, J. R. (2006). Computational prediction of methylation status in human genomic sequences. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(28), 10713–10716.

Dion, M. F., Altschuler, S. J., Wu, L. F., & Rando, O. J. (2005). Genomic characterization reveals a simple histone H4 acetylation code. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(15), 5501–5506.

Esteller, M. (2007). Cancer epigenomics: DNA methylomes and histone-modification maps. *Nature Reviews. Genetics*, *8*(4), 286–298.

Fan, S., Zhang, M. Q., & Zhang, X. (2008). Histone methylation marks play important roles in predicting the methylation status of CpG islands. *Biochemical and Biophysical Research Communications*, *374*(3), 559–564.

Fang, F., Fan, S., Zhang, X., & Zhang, M. Q. (2006). Predicting methylation status of CpG islands in the human brain. *Bioinformatics (Oxford, England)*, *22*(18), 2204–2209.

Feinberg, A., & Irizarry, R. (2010). Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proceedings of the National Academy of Sciences, Early Edition*.

Feltus, F. A., Lee, E. K., Costello, J. F., Plass, C., & Vertino, P. M. (2006). DNA motifs associated with aberrant CpG island methylation. *Genomics*, *87*(5), 572–579.

Field, Y., Fondufe-Mittendorf, Y., Moore, I. K., Mieczkowski, P., Kaplan, N., & Lubling, Y. (2009). Gene expression divergence in yeast is coupled to evolution of DNA-encoded nucleosome organization. *Nature Genetics*, *41*(4), 438–445.

Field, Y., Kaplan, N., Fondufe-Mittendorf, Y., Moore, I. K., Sharon, E., & Lubling, Y. (2008). Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS Computational Biology*, *4*(11), e1000216.

Francis, N. J., & Kingston, R. E. (2001). Mechanisms of transcriptional memory. *Nature Reviews. Molecular Cell Biology*, *2*(6), 409–421.

Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R., & Young, R. A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, *130*(1), 77–88.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., & Feldser, D. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, *458*(7235), 223–227.

Hauenschild, A., Ringrose, L., Altmutter, C., Paro, R., & Rehmsmeier, M. (2008). Evolutionary plasticity of polycomb/trithorax response elements in Drosophila species. *PLoS Biology*, *6*(10), e261.

Hawkins, R. D., Hon, G. C., & Ren, B. (2010). Next-generation genomics: an integrative approach. *Nature Reviews. Genetics*, *11*(7), 476–486.

Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., & Harp, L. F. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, *459*(7243), 108–112.

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., & Hawkins, R. D. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics*, *39*(3), 311–318.

Ioshikhes, I., Bolshoy, A., Derenshteyn, K., Borodovsky, M., & Trifonov, E. N. (1996). Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. *Journal of Molecular Biology*, *262*(2), 129–139.

Ioshikhes, I. P., Albert, I., Zanton, S. J., & Pugh, B. F. (2006). Nucleosome positions predicted through comparative genomics. *Nature Genetics*, *38*(10), 1210–1215.

Jenuwein, T., & Allis, C. D. (2001). Translating the histone code. *Science*, *293*(5532), 1074–1080.

Jiang, C., & Pugh, B. F. (2009). Nucleosome positioning and gene regulation: Advances through genomics. *Nature Reviews. Genetics*, *10*(3), 161–172.

Jirtle, R. L., & Skinner, M. K. (2007). Environmental epigenomics and disease susceptibility. *Nature Reviews. Genetics*, *8*(4), 253–262.

Johnson, S. M., Tan, F. J., McCullough, H. L., Riordan, D. P., & Fire, A. Z. (2006). Flexibility and constraint in the nucleosome core landscape of Caenorhabditis elegans chromatin. *Genome Research*, *16*(12), 1505–1516.

Jones, P. A., & Baylin, S. B. (2002). The fundamental role of epigenetic events in cancer. *Nature Reviews. Genetics*, *3*(6), 415–428.

Kaplan, N., Hughes, T. R., Lieb, J. D., Widom, J., & Segal, E. (2010, Nov 30). Contribution of histone sequence preferences to nucleosome organization: proposed definitions and methodology. *Genome Biology*, *11*(11), 140.

Kaplan, N., Moore, I. K., Fondufe-Mittendorf, Y., Gossett, A. J., Tillo, D., & Field, Y. (2009). The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, *458*(7236), 362–366.

Keshet, I., Schlesinger, Y., Farkash, S., Rand, E., Hecht, M., & Segal, E. (2006). Evidence for an instructive mechanism of de novo methylation in cancer cells. *Nature Genetics*, *38*(2), 149–153.

Kornberg, R. D. (1974). Chromatin structure: A repeating unit of histones and DNA. *Science*, *184*(139), 868–871.

Kornberg, R. D., & Lorch, Y. (1999). Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell*, *98*(3), 285–294.

Kornberg, R. D., & Stryer, L. (1988). Statistical distributions of nucleosomes: Nonrandom locations by a stochastic mechanism. *Nucleic Acids Research*, *16*(14A), 6677–6690.

Kouzarides, T. (2007). Chromatin modifications and their function. *Cell*, *128*(4), 693–705.

Ku, M., Koche, R. P., Rheinbay, E., Mendenhall, E. M., Endoh, M., & Mikkelsen, T. S. (2008). Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLOS Genetics*, *4*(10), e1000242.

Kurdistani, S. K., & Grunstein, M. (2003). Histone acetylation and deacetylation in yeast. *Nature Reviews. Molecular Cell Biology*, *4*(4), 276–284.

Lan, F., Nottke, A. C., & Shi, Y. (2008). Mechanisms involved in the regulation of histone lysine demethylases. *Current Opinion in Cell Biology*, *20*(3), 316–325.

Lantermann, A. B., Straub, T., Strålfors, A., Yuan, G. C., Ekwall, K., & Korber, P. (2010). Schizosaccharomyces pombe genome-wide nucleosome mapping reveals positioning mechanisms distinct from those of Saccharomyces cerevisiae. *Nature Structural & Molecular Biology*, *17*(2), 251–257.

Lee, J. T. (2009). Lessons from X-chromosome inactivation: Long ncRNA as guides and tethers to the epigenome. *Genes & Development*, *23*(16), 1831–1842.

Lee, W., Tillo, D., Bray, N., Morse, R. H., Davis, R. W., & Hughes, T. R. (2007). A high-resolution atlas of nucleosome occupancy in yeast. *Nature Genetics*, *39*(10), 1235–1244.

Lewis, E. B. (1978). A gene complex controlling segmentation in Drosophila. *Nature*, *276*(5688), 565–570.

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., & Tonti-Filippini, J. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, *462*(7271), 315–322.

Liu, Y., Shao, Z., & Yuan, G. C. (2010). Prediction of polycomb target genes in mouse embryonic stem cells. *Genomics*, *96*(1), 17–26.

Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F., & Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8. A resolution. *Nature*, *389*(6648), 251–260.

Martens, J. H., O'Sullivan, R. J., Braunschweig, U., Opravil, S., Radolf, M., & Steinlein, P. (2005). The profile of repeat-associated histone lysine methylation states in the mouse epigenome. *The EMBO Journal*, *24*(4), 800–812.

Mavrich, T. N., Ioshikhes, I. P., Venters, B. J., Jiang, C., Tomsho, L. P., & Qi, J. (2008). A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Research*, *18*(7), 1073–1083.

Mavrich, T. N., Jiang, C., Ioshikhes, I. P., Li, X., Venters, B. J., & Zanton, S. J. (2008). Nucleosome organization in the Drosophila genome. *Nature*, *453*(7193), 358–362.

Meissner, A., Mikkelsen, T.S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., et al. (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*.

Miele, V., Vaillant, C., d'Aubenton-Carafa, Y., Thermes, C., & Grange, T. (2008). DNA physical properties determine nucleosome occupancy from yeast to fly. *Nucleic Acids Research*, *36*(11), 3746–3756.

Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., & Giannoukos, G. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, *448*(7153), 553–560.

Moazed, D. (2009). Small RNAs in transcriptional gene silencing and genome defence. *Nature*, *457*(7228), 413–420.

Mohn, F., Weber, M., Rebhan, M., Roloff, T. C., Richter, J., & Stadler, M. B. (2008). Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. *Molecular Cell*, *30*(6), 755–766.

Morozov, A. V., Fortney, K., Gaykalova, D. A., Studitsky, V. M., Widom, J., & Siggia, E. D. (2009). Using DNA mechanics to predict in vitro nucleosome positions and formation energies. *Nucleic Acids Research*, *37*(14), 4707–4722.

Okita, K., Ichisaka, T., & Yamanaka, S. (2007). Generation of germline-competent induced pluripotent stem cells. *Nature*, *448*(7151), 313–317.

Ozsolak, F., Song, J. S., Liu, X. S., & Fisher, D. E. (2007). High-throughput mapping of the chromatin structure of human promoters. *Nature Biotechnology*, *25*(2), 244–248.

Peckham, H. E., Thurman, R. E., Fu, Y., Stamatoyannopoulos, J. A., Noble, W. S., & Struhl, K. (2007). Nucleosome positioning signals in genomic DNA. *Genome Research*, *17*(8), 1170–1177.

Pokholok, D. K., Harbison, C. T., Levine, S., Cole, M., Hannett, N. M., & Lee, T. I. (2005). Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell*, *122*(4), 517–527.

Rando, O. J., & Chang, H. Y. (2009). Genome-wide views of chromatin structure. *Annual Review of Biochemistry*, *78*, 245–271.

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., & Simon, I. (2000). Genome-wide location and function of DNA binding proteins. *Science*, *290*(5500), 2306–2309.

Richmond, T. J., & Davey, C. A. (2003). The structure of DNA in the nucleosome core. *Nature*, *423*(6936), 145–150.

Ringrose, L., Rehmsmeier, M., Dura, J. M., & Paro, R. (2003). Genome-wide prediction of Polycomb/Trithorax response elements in Drosophila melanogaster. *Developmental Cell*, *5*(5), 759–771.

Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., & Brugmann, S. A. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*, *129*(7), 1311–1323.

Robert, F., Pokholok, D. K., Hannett, N. M., Rinaldi, N. J., Chandy, M., & Rolfe, A. (2004). Global position and recruitment of HATs and HDACs in the yeast genome. *Molecular Cell*, *16*(2), 199–209.

Roh, T. Y., Cuddapah, S., & Zhao, K. (2005). Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes & Development*, *19*(5), 542–552.

Saccani, S., & Natoli, G. (2002). Dynamic changes in histone H3 Lys 9 methylation occurring at tightly regulated inducible inflammatory genes. *Genes & Development*, *16*(17), 2219–2224.

Satchwell, S. C., Drew, H. R., & Travers, A. A. (1986). Sequence periodicities in chicken nucleosome core DNA. *Journal of Molecular Biology*, *191*(4), 659–675.

Schlesinger, Y., Straussman, R., Keshet, I., Farkash, S., Hecht, M., & Zimmerman, J. (2007). Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. *Nature Genetics*, *39*(2), 232–236.

Schones, D. E., Cui, K., Cuddapah, S., Roh, T. Y., Barski, A., & Wang, Z. (2008). Dynamic regulation of nucleosome positioning in the human genome. *Cell*, *132*(5), 887–898.

Schuettengruber, B., Chourrout, D., Vervoort, M., Leblanc, B., & Cavalli, G. (2007). Genome regulation by polycomb and trithorax proteins. *Cell*, *128*(4), 735–745.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thastrom, A., Field, Y., & Moore, I. K. (2006). A genomic code for nucleosome positioning. *Nature*, *442*(7104), 772–778.

Segal, E., & Widom, J. (2009). What controls nucleosome positions? *Trends in Genetics*, *25*(8), 335–343.

Sekinger, E. A., Moqtaderi, Z., & Struhl, K. (2005). Intrinsic histone-DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast. *Molecular Cell*, *18*(6), 735–748.

Seligson, D. B., Horvath, S., Shi, T., Yu, H., Tze, S., & Grunstein, M. (2005). Global histone modification patterns predict risk of prostate cancer recurrence. *Nature*, *435*(7046), 1262–1266.

Sharma, S., Kelly, T. K., & Jones, P. A. (2010). Epigenetics in cancer. *Carcinogenesis*, *31*(1), 27–36.

Simon, J. A., & Kingston, R. E. (2009). Mechanisms of polycomb gene silencing: Knowns and unknowns. *Nature Reviews. Molecular Cell Biology*, *10*(10), 697–708.

Sing, A., Pannell, D., Karaiskakis, A., Sturgeon, K., Djabali, M., & Ellis, J. (2009). A vertebrate Polycomb response element governs segmentation of the posterior hindbrain. *Cell*, *138*(5), 885–897.

Sparmann, A., & van Lohuizen, M. (2006). Polycomb silencers control cell fate, development and cancer. *Nature Reviews. Cancer*, *6*(11), 846–856.

Strahl, B. D., & Allis, C. D. (2000). The language of covalent histone modifications. *Nature*, *403*(6765), 41–45.

Svaren, J., & Horz, W. (1997). Transcription factors vs nucleosomes: Regulation of the PHO5 promoter in yeast. *Trends in Biochemical Sciences*, *22*(3), 93–97.

Takahashi, K., & Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, *126*(4), 663–676.

TCGA. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, *455*(7216), 1061–1068.

Tirosh, I., Berman, J., & Barkai, N. (2007). The pattern and evolution of yeast promoter bendability. *Trends in Genetics*, *23*(7), 318–321.

Waddington, C. (1942). The epigenotype. *Endeavour*, *1*, 18–20.

Wang, A., Kurdistani, S. K., & Grunstein, M. (2002). Requirement of Hos2 histone deacetylase for gene activity in yeast. *Science*, *298*(5597), 1412–1414.

Wang, L., Brown, J. L., Cao, R., Zhang, Y., Kassis, J. A., & Jones, R. S. (2004). Hierarchical recruitment of polycomb group silencing complexes. *Molecular Cell*, *14*(5), 637–646.

Wei, G., Wei, L., Zhu, J., Zang, C., Hu-Li, J., & Yao, Z. (2009). Global mapping of H3K4me3 and H3K27me3 reveals specificity and plasticity in lineage fate determination of differentiating CD4+ T cells. *Immunity*, *30*(1), 155–167.

Wernig, M., Meissner, A., Foreman, R., Brambrink, T., Ku, M., & Hochedlinger, K. (2007). In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. *Nature*, *448*(7151), 318–324.

Widom, J. (2001). Role of DNA sequence in nucleosome stability and dynamics. *Quarterly Reviews of Biophysics*, *34*(3), 269–324.

Woo, C. J., Kharchenko, P. V., Daheron, L., Park, P. J., & Kingston, R. E. (2010). A region of the human HOXD cluster that confers polycomb-group responsiveness. *Cell*, *140*(1), 99–110.

Yoo, C. B., & Jones, P. A. (2006). Epigenetic therapy of cancer: Past, present and future. *Nature Reviews. Drug Discovery*, *5*(1), 37–50.

Yu, J., Vodyanik, M. A., Smuga-Otto, K., Antosiewicz-Bourget, J., Frane, J. L., & Tian, S. (2007). Induced pluripotent stem cell lines derived from human somatic cells. *Science*, *318*(5858), 1917–1920.

Yuan, G. C. (2009). Targeted recruitment of histone modifications in humans predicted by genomic sequences. *Journal of Computational Biology*, *16*(2), 341–355.

Yuan, G. C., & Liu, J. S. (2008). Genomic sequence is highly predictive of local nucleosome depletion. *PLoS Computational Biology*, *4*(1), e13.

Yuan, G. C., Liu, Y. J., Dion, M. F., Slack, M. D., Wu, L. F., & Altschuler, S. J. (2005). Genome-scale identification of nucleosome positions in S. cerevisiae. *Science*, *309*(5734), 626–630.

Zhang, Y., Moqtaderi, Z., Rattner, B. P., Euskirchen, G., Snyder, M., & Kadonaga, J. T. (2009). Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. *Nature Structural & Molecular Biology*, *16*(8), 847–852.

Zhao, J., Sun, B. K., Erwin, J. A., Song, J. J., & Lee, J. T. (2008). Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science*, *322*(5902), 750–756.

Zhou, V. W., Goren, A., & Bernstein, B. E. (2011). Charting histone modifications and the functional organization of mammalian genomes. *Nature Reviews. Genetics*, *12*(1), 7–18.

KEY TERMS AND DEFINITONS

Epigenetics: The study of inherited changes in phenotype or gene expression caused by mechanisms other than changes in the underlying DNA sequence.

Nucleosome: The basic unit of DNA packaging in eukaryotes, consisting of a segment of DNA wound around a histone protein core.