Stony Brook University

Epigenetic Study with Genome-wide Hypothesis Test and

Stepwise Multivariate Adaptive Regression Splines (SMARS)

A Dissertation Presented

by

Yijin Wu

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

December 2014

Stony Brook University

The Graduate School

Yijin Wu

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Song Wu – Dissertation Advisor
Assistant Professor, Department of Applied Mathematics and Statistics

Wei Zhu – Dissertation Co-advisor
Professor, Deputy Chair, Department of Applied Mathematics and Statistics

Jie Yang - Chairperson of Defense
Assistant Professor, Department of Preventive Medicine, Director, Statistical Consulting Core, School of Medicine, Adjunct Assistant Professor, Department of Applied Mathematics and Statistics

Colette Pameijer – Outside Member
Associate Professor of Surgery, Division of General Surgery and Surgical Oncology
Penn State Hershey Medical Center

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

Epigenetic Study with Genome-wide Hypothesis Test and Stepwise Multivariate Adaptive Regression Splines (SMARS)

by

Yijin Wu

Doctor of Philosophy

in

Department of Applied Mathematics and Statistics

Stony Brook University

2014

Epigenetic gene regulations are essential processes for development and differentiation in both animals and plants. With the advent and rapid advance of sequencing techniques, the high-throughput genome-wide epigenetic modification profiles have been extensively studied in the past few years. In this thesis work, we studied the relationship between gene regulation and two major epigenetic modifications, i.e., DNA methylation and histone modifications.

In the DNA methylation analysis, we studied two strains of Arabidopsis grown under different levels of carbon dioxide concentrations (430ppm vs. 810ppm) to simulate the impact of global climate change. The differentially methylated regions were identified by genome-wide hypothesis tests and the potentially impacted genes were located on the genome. We successfully detected the differentially expressed genes that function in plants development. This study illustrated how plants adapted to the environmental stress through epigenetic mechanism.

In histone modification analysis, we proposed a data-driven model developed from Multivariate Adaptive Regression Splines (MARS). This step-wise MARS model is able to capture interactions among different chromatin features as well as among genomic loci. Not only can our method outperform existing methods in terms of prediction accuracy, it can also identify potential interactions that could shed light on further study of histone code hypothesis.

Table of Contents

List of Tables

List of Figures

Acknowledgment

I wish to express my sincere thanks to Department of Applied Mathematics and Statistics, for providing me with all the necessary facilities.

I would like to thank my advisor Professor Song Wu and co-advisor Professor Wei Zhu for advising me on my algorithm development work in thesis. I also appreciate their support on my research assistant work. I would also like to thank Dr. Qiong Liu (Alison) in the department of Biochemistry and Cell Biology, who introduced me to epigenetics study and advised on methylation study. Lastly I would also thank Dr. Colette Pameijer for giving me chance to work with her on oncology study.

Chapter I: Introduction

1 Epigenetic regulation

Epigenetic regulation of gene process is essential for development and differentiation in both animals and plants. It refers to functionally relevant modifications to genomes that do not involve any changes in nucleotide sequences. The epigenetic status of one whole genome is termed as epigenome. The epigenome is usually very dynamic, and can establish and maintain cell type-specific gene expression profile that features cellular identity and function. Studies have indicated epigenetic modifications of chromatin and its response to distinct environmental factors may directly contribute to developmental processes (Schones et al., 2008). Typically there are two major epigenetics mechanisms: one is DNA methylation and the other is histone modification (Figure 1).

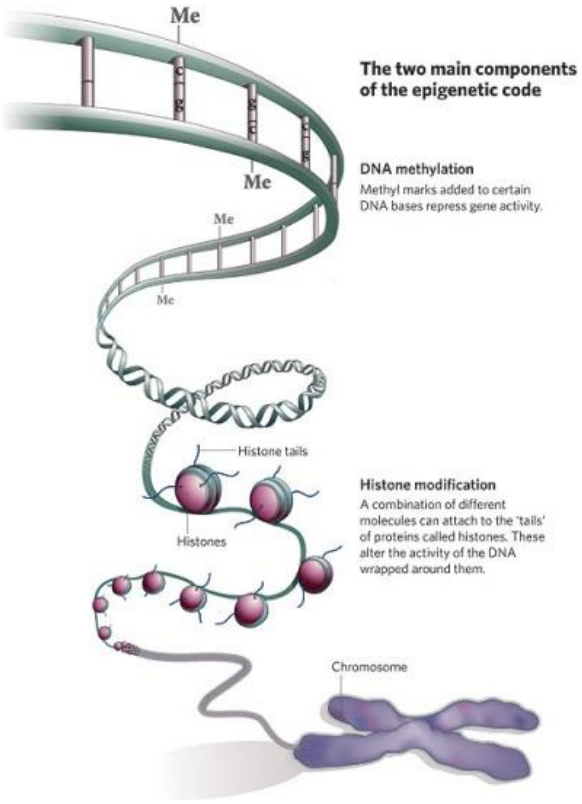


Figure 1. An illustration of epigenetic mechanisms (Reynolds et al., 2012). Two main components of epigenetic modification are DNA methylation and histone modification. Both are functional in gene regulation. DNA methylation is the addition of methyl group to the cytosine or ardenine DNA nucleotides. A large number of histone modifications are histone methylation or acetylation, which are the addition of methyl group or acetyl group to the N-tail of the histones.

The history of epigenetics study is linked to the study of evolution and development. It can be traced back to the earlier studies in the 19th century when scientists began to present the understanding between genes and development. The initial definition of epigenetics was very vague. It encompassed almost all regulated processes that shape the final product with genetic materials.

There was a long debate on how a single fertilized egg can rise to a complex organism with cells of varied phenotypes. Another case was how different types of blood cells can be generated from the common bone marrow stem cells. There clearly existed switches in the gene activity related to cell differentiations. The chromosome X was an example of such switch mechanism. At the early stage of development, one X chromosome was randomly inactivated in every cell while the other was activated. As the two X chromosomes contain almost identical DNA sequences, it seemed that the inactivity and activity of chromosome was intrinsic to themselves. This not only suggested a switch mechanism in the early development but its subsequent heritability. Later, many studies provided strong evidence that the developmental program did reside in the chromosomes. The important role of DNA methylation proposed in 1969 provided the basic outline of the switch for gene activities during development. Riggs and collegues (Riggs 1975) addressed the methylation mechanism of chromosome X inactivation. Studies also indicated that such pattern of methylation could be heritable if a specific enzyme existed.

With the cloning and sequencing of DNA, methods were developed to screen DNA methylation in specific DNA sequences. It was soon discovered that many genes with methylation at the promoters are inactive. Originally, it was thought that the variability of gene expression was due to mutations, but now it has become apparent that the aberrant changes in the distribution of 5-methyl cytosine could also result in changes in gene expression. Ultimately, the term “epigenetics” was redefined so as to distinguish heritable changes that arose from sequence changes in DNA from those that did not.

In the past years, there has also been a great interest in examining the association between histone modification and gene expression. Well before most of the work on DNA methylation,

Stedman (Stedman et al., 1950) claimed that the histones could act as general repressors of gene expression. Followed by studies addressing the capacity of chromatin to serve as a template for transcriptional activities and the discovery that DNA was packaged into the nucleosome, it is much easier to study how histone modification may affect gene expression. Nucleosome is the fundamental histone-containing chromatin subunit. The chromatin structure and gene expression has become a very active area of study.

Studies in animal models have demonstrated that epigenomic variability leads to phenotypic variability, such as disease susceptibility that is not recognized at the DNA sequence level. Epigenetic processes may respond to various external factors including environment, diet, and even behavior. This plasticity is thought to allow the organism to respond and adapt quickly to external stress, yet also confer the organism, and even in some cases its offspring, with the ability to "memorize" contacts with such stress into adulthood (Dolinoy et al., 2008; Morgan et al., 2008). Although many studies have been conducted on how epigenomic variability may contribute to disease susceptibility in humans, this is still a largely unexplored area. Among those studies relevant to human diseases, in particular in cancer, substantial alternations of DNA methylation and histone modifications have been described. Global hypomethylation and site-specific gene hypermethylation of DNA are so widespread that they are now considered hallmarks of cancer (Feinberg et al., 2004). Environmental exposure to carcinogens has also been linked to these alternations (Sutherland et al., 2003; Egger et al., 2004). More importantly, some of such epigenetic alternations have been shown to be inheritable through the germline (Suter et al., 2004; Chan et al., 2006).

Although today we have some ideas of the establishment and maintenance of the epigenetic modifications, the true relationship between gene regulation and its mechanism is still largely unknown.

2 DNA methylation

DNA methylation refers to the addition of a methyl group to the 5th position of the cytosine pyrimidine ring (Figure 2).

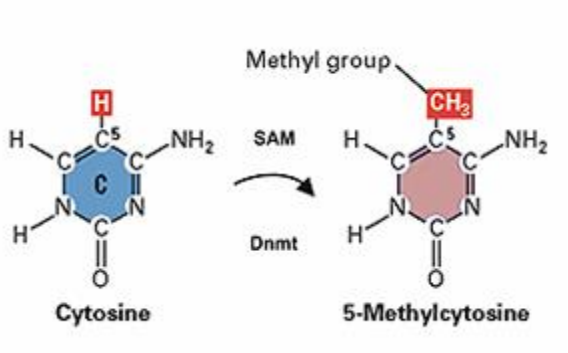


Figure 2. The chemistry modification of DNA methylation. DNA methylation occurs when cytosine bases are converted to 5-methylcytosine by DNA methyltransferase (DNMT) enzymes. (http://www.hgu.mrc.ac.uk/people/r.meehan_researchb.html)

Prior to 1980, there had been a number of clues suggesting that methylation might play a role in the regulation of gene expression. Shortly after McGhee and Ginder published their discovering that DNA methylation may be involved in gene expression (McGhee et al., 1979), direct experiments were performed to examine the inhibiting effect of methylation on gene expression (Jones et al., 1980). By comparing the cells before and after 5-azacytidine (one of chemical analogs for the nucleoside cytidine) treatment, the straightforward experiments

demonstrated that the analogs that impact methylation were responsible for the cellular differentiation.

It is known today that the DNA methylation happens when cytosine bases are converted to 5-methylcytosine by DNA methyltransferase (DNMT) enzymes. Different enzymes in DNMT family either act as de novo DNMTs by adding the initial pattern of methyl groups on DNA sequence, or as maintenance by copying the methylation from an existing DNA strand after replication. In mammals, DNA methylation patterns are established by the DNA methyltransferase 3 (DNMR3) (Goll et al., 2005; Cheng et al., 2008; Kim et al., 2009). While in plants, the de novo DNA methylation is catalyzed by Domain Rearraned Methytransferase 2 (DRM2), a homologue of the DNMT3. The plants' methylation is maintained by three different pathways: CG methylation is maintained by DNA methyltransferase 1 (MET1); CHG methylation by chromomrhylase 3 (CMT3, plant-specific DNA methyltransferase); and symmetric CHH methylation by DRM2 through persistent de novo methylation.

Besides the establishment and maintenance, the targets and patterns of DNA methylation have also drawn a lot of attention. Mammals tend to have sparsely but globally distributed CpG methylation throughout the entire genome, except for CpG islands or sequences with high contents of CpG. The heavily methylated regions are found interspersed with un-methylated regions (Figure 3c). This "mosaic" global pattern of DNA methylation in mammals makes it difficult to determine genomic targets of methylation. In plants, up to 50% of their cytosine residues exhibiting methylation due to large number of transposons (Figure 3e). DNA methylation could occur on cytosine in more flexible sequence contexts, including the symmetrical CpG, CHG sequences and asymmetric CHH sequences (H = A,T or C) . While in fungi, only the repetitive DNA sequences are found methylated (Figure 3a). Although plants like

Arabidopsis show similar mosaic methylation patterns to those of animals (Figure 3b), it has been indicated that DNA methylation are different between animals and plants. One significant difference is that in some plants, methylation is absent or occurs altogether on transposable elements, and this mechanism has been revealed to involve small interfering RNA (siRNA) (Mette et al., 2000; Chan et al., 2004; Chan et al., 2005).

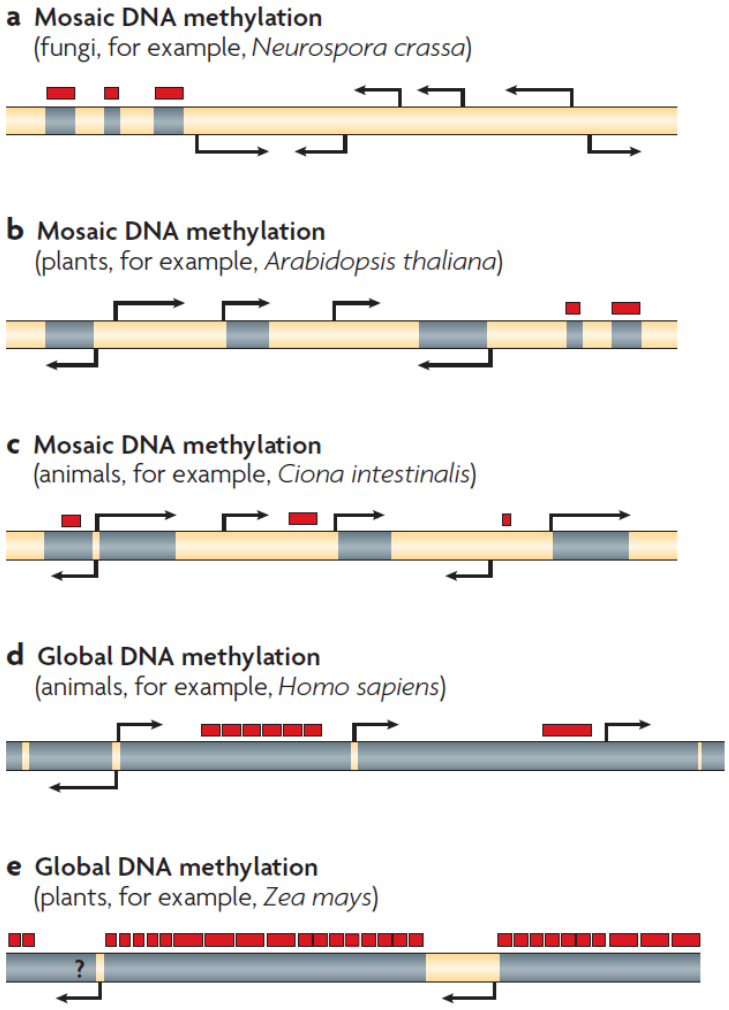


Figure 3. DNA methylation landscapes in fungi, animals and plants (Suzuki et al., 2008). The grey box and yellow box denote the stable methylated and unmethylated domains, respectively. The red box indicates the transposable elements. (a) DNA methylation in fungi is interspersed by methylated and unmethylated domains. In certain fungi the efficient targeted methylation was

observed at transposable elements. (b) The plant Arabidopsis thaliana illustrates a mosaic methylation pattern that is due to gene-body methylation. Unlike animals, transposons and repetitive elements are subject to targeted RNA-mediated methylation. (c) Mosaic methylation is also characteristic of most invertebrates, but has only been mapped in detail in the sea squirt Ciona intestinalis. Gene-body methylation affects over half of all genes, but the remainder is embedded within unmethylated DNA. Transposable elements are frequently unmethylated. (d) Vertebrate genomes are globally methylated, with only CpG islands being unmethylated. Transposable elements are methylated, as are gene bodies and intergenic regions. (e) The DNA methylation landscape of plants with large genomes, such as maize, has not been mapped in detail, but it is evident that genes are separated by long tracts of DNA that contain transposable elements. Genes tend to be unmethylated, but the existence of gene-body targeted methylation has not yet been discussed.

Much research has suggested that, cytosines can silence or activate the gene expression by gaining or losing methylation at particular sites. DNA methylation of promoter regions usually inhibits transcription by blocking the binding of transcription factors with the promoters, while in coding regions, it generally does not affect gene expression (Wolffe et al., 1999; Urnov et al., 2001).

Functionally, DNA methylation has been linked to many important biological processes. Its link to cancer was first reported in 1983, where it was shown that genomes of cancer cells were hypomethylated due to the loss of methylation from repetitive genome regions. (Feinberg et al., 2004). The hypomethylation caused genomic instability which was a hallmark of tumor cells. In the studies of plants, decreased methylation also led to a number of phenotypic and developmental abnormalities (Finnegan et al., 1996).

The distribution of methylation marks could convey the enormous volume of epigenetic information of transcriptional repression or activation. Therefore, the broader DNA methylation profiles are important implications for understanding why certain genes can be expressed under specific contexts and how epigenetic changes might be related to abnormality or disease.

Although evidence indicates the role of DNA methylation in repressing gene expression especially in promoters, the exact function of methylation is largely uncovered. However, the overall similarity and differences of DNA methylation levels within and between eukaryotic groups suggested that they might share a common underlying mechanism. In order to show this, a thorough understanding of genome distribution of the cytosine methylation of a variety of species is needed.

So far, various high-throughput approaches have been developed and applied to the genome-wide analysis. Novel techniques such as the Next Generation Sequencing (NGS) can provide single base pair resolution, quantifying DNA methylation with genome wide coverage. There are three major approaches distinguishing the methylated and un-methylated cytosines, including methylation sensitive restriction enzyme digestion, affinity purification and bisulfite conversion of DNA. Any of these methods can be combined with either hybridization to DNA microarrays or direct sequencing to study DNA methylation on a genomic scale.

Restriction enzyme DNA methylation method can identify methylated sites using restriction enzymes that differentially recognize methylated and unmethylated cytosine bases. The enzymes (and recognition sites) include AciI (CCGC and GCGG), BstUI (CGCG), HhaI (GCGC) and TaiI (ACGT). For the affinity purification methods, DNA is first sonicated and a methyl-cytosines specific antibody is used to pull down methylated regions. Besides these two approaches, the bisulfite treated DNA methylation sequencing is also widely used.

Bisulfite conversion method is based on the selective deamination of cytosine, but not of 5-methylcytosine, with the treatment of sodium bisulfite (Figure 4). If the genomic DNA were treated with sodium bisulfite, all the un-methylated cytosine would be converted to uracil, which

is subsequently amplified as thymine during polymerase chain reaction (PCR). On the other hand, the methylated cytosines would not be affected and remain cytosines during PCR. Hence, in the final sequencing, the 5-methycytosine will be still "read" as cytosine, but unmethylated cytosine becomes thymine. By comparing sequencing data before and after the sodium bisulfite treatment, the methylated cytosine sites can then be identified. The sequencing from bisulfite converted DNA usually provides very reliable and detailed information on the methylation pattern for every single cytosine, and has been regarded as the "gold" standard of DNA methylation analysis (Laird 2010, Zhang and Jeltsch 2010).

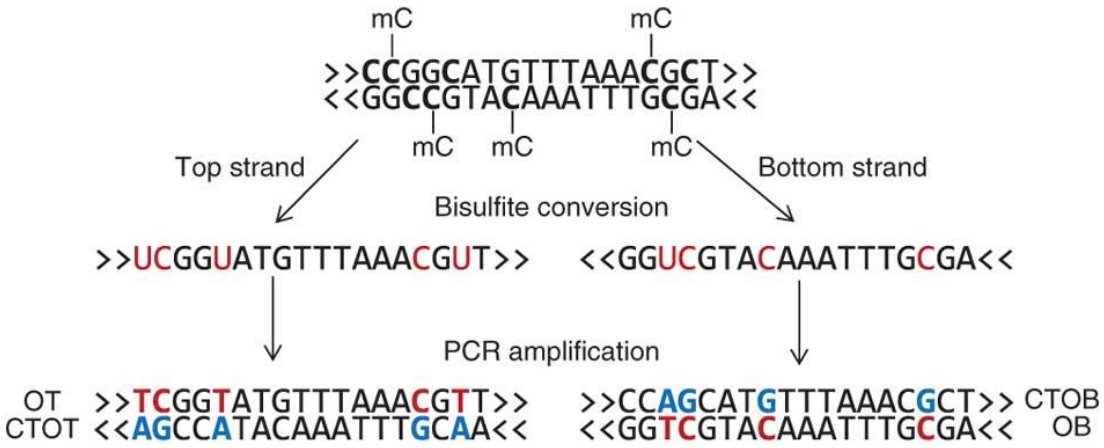


Figure 4. Bisulfite conversion and PCR amplification. Bisulfite conversion of genomic DNA and subsequent PCR amplification gives rise to two PCR products and up to four potentially different DNA fragments for any given locus(Krueger et al., 2012). Methylated cytosine residues are resistant to bisulfite conversion and are used to detect the DNA methylation state. mC, 5-methylcytosine; OT, original top strand; CTOT, strand complementary to the original top strand; OB, original bottom strand; and CTOB, strand complementary to the original bottom strand.

3 Histone modifications

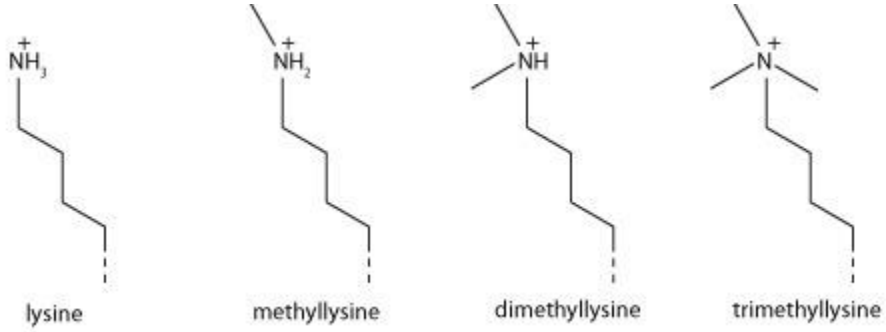
Histone modification is another important epigenetic mechanism in gene regulation. There are five major families of histones: H1/H5, H2A, H2B, H3 and H4. Histones H2A, H2B, H3, and H4 are four core histones in higher organisms, while H1 and H5 are known as the linker histones. Two copies each of the core histones consist of a histone octamer, around which ~146 base pairs of DNA can be wrapped to form a nucleosome. The nucleosome is the fundamental subunit of chromatin. As histones act cooperatively to prepare the chromatin for transcriptional activations, it is another important mechanism for the epigenetic regulation.

Identification of the enzymes for histone modifications has been the focus of intense research in the past decade, including histone acetylation (Sterner et al., 2000), methylation(Zhang et al., 2001), phosphorylation(Nowak et al., 2004), ubiquitination(Shilatifard 2006), sumoylation(Nathan et al., 2006), ADP-ribosylation(Hassa et al., 2006), deamination(Cuthbert et al., 2004; Wang et al., 2004), and proline isomerization(Nelson et al., 2006). Among all the enzymes, the methyltransferases and kinases are more specific characterized ones so far.

In addition, histone methylation and acetylation of specific lysine residues on the N-terminal histone tails are fundamental for the formation of chromatin domains. Histone methyltransferases are enzymes which transfer methyl groups from S-Adenosyl methionine onto the lysine or arginine residues of the histones. The well-known lysine methylations include mono-, di-, or tri-methylation (Figure 5a). Different degrees of residue methylation can

havedifferent functions. For example, mono-methylated H4k20 (H4k20me1) is involved in transcriptional repression while tri-methylated H4k20 (H4k20me3) serves in chromatin repression. Histone acetylation (Figure 5b) refers to the acetylation of histone tails induced by histone acetyltransferase enzyme family (HATs). Currently histone acetylation is thought to be associated with transcription activation. The terminology of histone modifications denotes the type of modification and the position of it taking place. For example, the H2k20me1 means the mono-methylation happens at the 20th Lysine amino acid at the histone H2 tail.

(a)



(b)

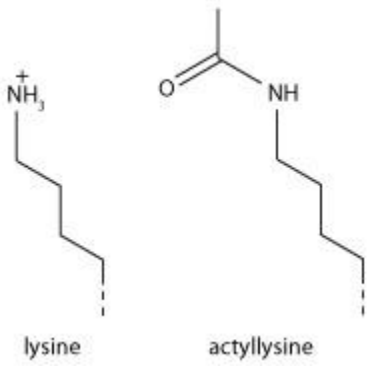


Figure 5. Biochemistry modifications on Lysine of histone tails (Wikipedia). (a) Histone methylation is addition of one, two, or three methyl groups to the N-terminals. Three different degree of methylation refer to mono-methylation (methylation), di-methylation and tri-methylation. (b) Histone acetylation is addition of acetyl group to the N-terminals.

Three models have been proposed to explain the function of histone modifications in gene regulation: the charge-neutralization model, the histone-code model and the signaling-pathway model (Schones et al., 2008). The function of histone acetylation was involved in the charge-neutralization model. It has been indicated that the histone acetylation can lead to a de-condensation of the chromatin structure(Wolffe et al., 1999; Turner 2000). With the relaxed structure, the specific regions of DNA are transcriptionally active. In this model, the histone modifications are responsible to the assortment of the genome into the transcriptionally active and transcriptionally slilent domains. The conept "histone code" was initially introduced to explain how a set of histone modificatios at the same genome region would control the gene regulation. The histone code hypothesis states that i) distinct histone modifications may induce interaction affinities for chromatin related proteins, ii) mofidications on the same or different histon tails may be inter-related and function combinationally, and iii) distinct qualities of higher order chromatin are highly dependent on the local concentration and combination of differentially modified nucleosomes (Jenuwein et al., 2001). The signalling pathway model is more general than the histone code model. It suggests that histone modifications could serve as signalling platforms to facilitate the binding of enzymes on the chromatin.

In the past few years, several technologies have been developed to explore the genome-wide histone modifications. Accommodated by development and improvement of the "ChIP-chip" technique, i.e., chromatin immunoprecipitation (ChIP) followed by the DNA-microarray analysis (chip), histone modification patterns have been extensively studied. The first studies with ChIP-chip suggested that histone modifications are associated with distinct genome regions and transcription states (Bernstein et al., 2002; Robyr et al., 2002; Schübeler et al., 2004). This

technique was also used to profile histone modifications in mammalian genomes. However, ChIP-chip has several limitations, like large sets of arrays needed to cover the mammalian genome and it can only detect targets included on the array.

ChIP can also be combined with the serial analysis of gene expression (SAGE) in a method termed genome-wide mapping technique (GMAT) to map several H3 modifications in human T cells. But this technology was still limited by a relatively low resolution and considerably high cost (Chiang et al., 2001). To overcome these problems, a method termed "ChIP-Seq" has been developed to employ the NGS technologies, such as Illumina Genome Analyzer, to directly sequence ChIP DNA fragments and infer the histone modifications along the genome.

The ChIP-Seq sequencing workflow includes two steps (Figure 6). At the ChIP step, the specific cross-linked DNA was processed by antibody against the protein of interest. Then the oligonucleotide adaptors are added to the small stretches of DNA. Short read sequences for the DNA templates are then sequenced simultaneously using a genome sequencer. The number of sequenced reads that are aligned to the reference genome is directly proportional to its modification level. Compared to the ChIP-chip, ChIP-seq is more quantitative and easier to be compared in terms of the modification levels.

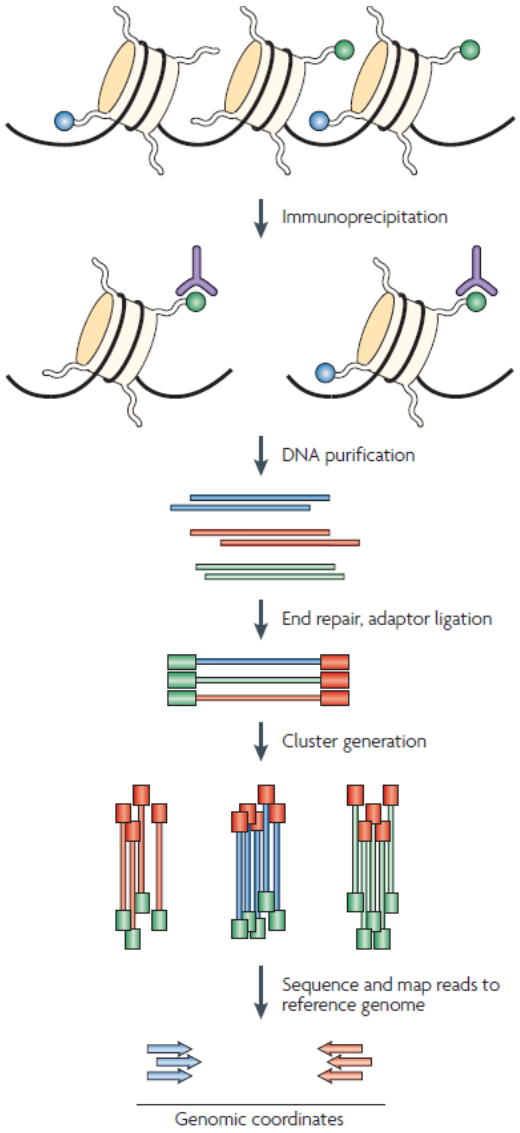


Figure 6. Workflow of ChIP-Sequencing for histone modification. The ChIP process enriches crosslinked DNA-protein using a protein specific antibody. After purify the DNA fragments, the oligonucleotide adapters are added to the fragments. Followed by a PCR cluster generation, the amplified clusters are directly sequenced. Then after alignment to the reference genome the mapped information could be used in the downstream data analysis. (Schones et al., 2008)

High-resolution profiles of the genome-wide distribution of histone lysine and arginine methylations, as well as their combinational patterns, have been reported (Chiang, Liu et al. 2001). It revealed that histone methylation patterns at promoters, enhancers and transcribed regions might be linked to gene activation or repression. Based on these, the positioning of

histone modification became another important area of epigenetic study. Several large-scale studies have shown that generally, high levels of histone acetylation and H3K4 methylation in the promoter regions of lead to the activation of gene expression (Bernstein, Kamal et al. 2005, Roh, Cuddapah et al. 2005, Roh, Cuddapah et al. 2006), whereas elevated levels of H3K27 methylation at promoters correlates with gene repression (Boyer, Plath et al. 2006, Lee, Jenner et al. 2006, Roh, Cuddapah et al. 2006).

With the genome-wide profiles of epigenetic modifications, we are gaining massive information of the DNA methylation and histone modifications. However, there are still a lot of unsolved problems regarding the functions of these two mechanisms.

In this thesis, I will study the association between epigenetic mechanisms with the gene regulation. In the second chapter, the impact of DNA methylation variation on adjacent gene expression which involved in plants development will be described. In the third chapter, the comprehensive correlation between histone modification signals will be approximated by a quantitative model.

Chapter II: Genome-wide DNA methylation variation study of *Arabidopsis thaliana*

In this chapter I will study the impact of DNA methylation variation on plant development. Specifically this is to identify the functional genes that may differentially express through the underlying stress-induced epigenetic modification.

1 Previous studies of DNA methylation variations

DNA methylation sites do not occur randomly, i.e., the clusters that distributed in specific regions are noteworthy. In the model plant *Arabidopsis thaliana*, there are about 24%, 6.7%, and 1.7% of cytosine methylation at CG, CHG, and CHH nucleotides, respectively (Cokus et al., 2008; Lister et al., 2008). In DNA methylation studies for *A. thaliana*, the genome-wide map of DNA methylation has been reported, including the genebody-specific methylation. It was found that about 18.9 % of the whole genome were signifiacnly methylated. (Zhang et al., 2006; Zilberman et al., 2007), particularly, high-frequency DNA methylations were found in pericentromeric heterochromatin, repetitive sequences and regions producing siRNAs. CHG methylation appeared mostly in the pericentromeric regions, due to its preference for methylation of transposon-related sequences (Tompa et al., 2002; Kato et al., 2003). In contrast to CHG, methylation of CG and CHH contexts, although mostly enriched in the pericentromeric regions, is commonly spread throughout the euchromatic chromosome arms.

Since the first study proposing that the cytosine methylation may influence gene expression (Holliday et al., 1975; Riggs 1975), it has been shown that frequency and patterns of methylation may be inherited, that DNA methylation can repress trancription, and that changes in DNA methylation are correlated with differences in gene expression in a tissue-specific manner. However, it is still unclear how DNA methylation may function as an essential role in regulating gene expression during development. In the past few years, DNA methylation studies aimed to profile the genome-wide distribution of the DNA methylation sites. Accompanied with high-throughput sequencing, the DNA methylation patterns of plant, as well as in mouse and human genomes, were studied. In higher eukaryotes, DNA methylation were found to be

involved in genomic imprinting and tumorigenesis in mammals, and in transposon silencing and gene regulation in plants (Li et al., 1992; Bestor 2000; Rhee et al., 2002; Lippman et al., 2004; Zhang et al., 2006).

Lots of evidence has shown that plant genes are both transcriptionally inactive and methylated within the promoters or coding regions. However, it is not clear that if the DNA methylation is the cause of gene inactivation or simply a consequence induced by other transcription factors. Although it is hard to conclude whether the DNA methylation is the cause or effect, studying the genome-wide distribution of DNA methylation and its correlation to functional genes would still help us explore the underlying epigenetic mechanism.

The methylation changes during development are very important. For plants, the loss of DNA methylation would induce abnormalities such as loss of apical dominance, reduced stature, altered leaf size and shape, reduced root length, homeotic transformation of floral organs, and reduced fertility (Kakutani et al., 1995; Finnegan et al., 1996; Ronemus et al., 1996). A recent stress-induced DNA methylation study found that specific stresses can trigger specific methylation alternations and therefore leading to epigenetic divergence(Verhoeven et al., 2010).

Recent studies indicated the potential inheritability of epigenetic change across generations. In study comparing ancestor and its 31^{th} generation under natural conditions, it was shown that, at the whole genome level, DNA methylation was very stable and heritable. Especially, the methylation of transposable elements was the most stable and consistent across the generations. (Becker et al., 2011; Schmitz et al., 2011). Some region that was found de-methylated in the 31^{th} generation was re-methylated in the following generation. This suggested that the DNA methylation may fluctuate within a short time and indicated existence of recurrent

cycles of forward and reverse mutations over long term. With the potential inheritability, the adaption of gene regulation to environmental impacts through the epigenetic regulation becomes extensively crucial in plants and mammals development. It was discussed that altered DNA methylation patterns were transmitted to the non-stressed progeny and may play a distinct role in stress-induced epigenetic inheritance in species evolution (Verhoeven et al., 2010).

Through the underlying epigenetic mechanism, enviromental factors may have influence on functional genes. Abnormallity in DNA methylation were found to related to development of disease in mammal such as cancer. As a major approach to study the plant reaction to the external factors, changes in DNA methylation induced by stress have drawn a lot of attention. It was indicated that DNA methylation may regulate plant development through the molecular basis of vernalization (Burn et al., 1993; Finnegan et al., 1998). This was the first indication that DNA methylation may regulate plant development. The author proposed that the vernalization is mediated by demethylation in the promoters of genes whose expression is functional for initiation of flowering. Decreased methylation was observed with a cold treatment. This might be the reason for early flowering with exposure to low temperature.

It was believd that the epigenomic alternations are the key to to adaption to climate changes. Study has already showed that the increasing temperatures affected the sexual plant reproductive phase (Hedhly et al., 2009). Other than temperature, stresses like drought, salt and pathogens were discussed in plants epigenetics study as well. These studies provided the insight of the plants spontaneous epimutation under regional conditions. However, as a major component in the photosynthetic reaction, the effect from different concentrations of carbon dioxide (CO_2) has not been discovered yet.

2 Main goal and significance of the study

Since the beginning of the Industrial Revolution, humans began burning coal in large quantities. With increased use of fossil fuels, carbon dioxide emission increased dramatically in the past century. The phenomenon of the so-called “global warming” is a direct consequence of increasing in carbon dioxide (CO2) in the Earth’s atmosphere. Prior to industrialization, the concentration of carbon dioxide in the atmosphere was 280 parts per million (ppm). The concentration is now approaching 500 ppm and the growth rate is still accelerating. According to the Intergovernmental Panel on Climate Change (IPCC), it is predicted that atmospheric CO2 levels could reach 500 ppm by 2050 and 800 ppm or more (Figure 7).

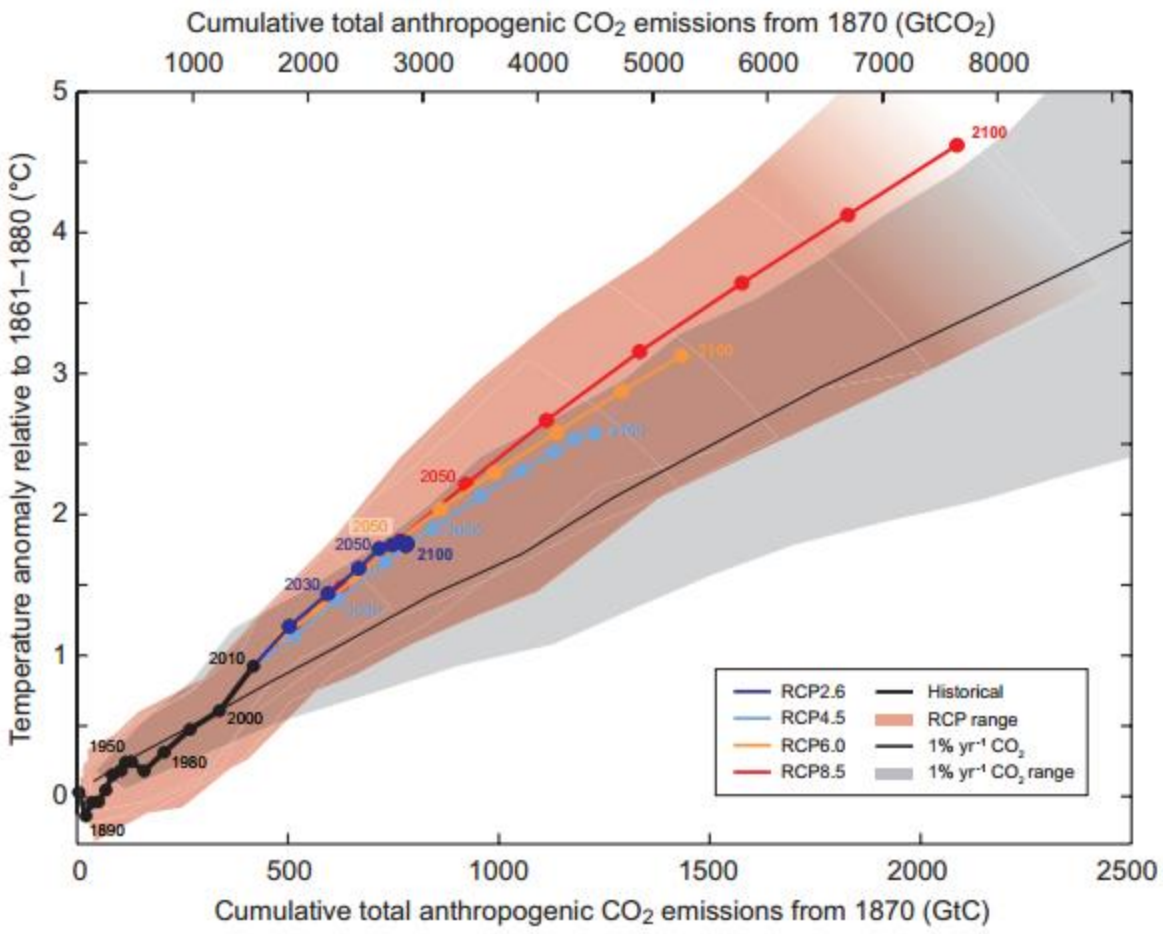


Figure 7. Global mean surface temperature increase as a function of cumulative total global CO2 emissions from various lines of evidence. In year 2050, we are likely to have double CO2 emission than we have now. RCPs, RCP2.6, RCP4.5, RCP6, and RCP8.5 are four possible climate futures depending on how much greenhouse gases are emitted in the years to come (IPCC Fifth Assessment Report, 2013)

The increasing of greenhouse gas in the atmosphere is changing the global environment greatly for mammals and plants, particularly for those economically important crops. With the explosion of population, the potentially reduced production of crops and fruits poses a serious threat to meet the increasing demand. Unlike the mammals which can migrant, plants have to adapt to the climate changes by self-regulating mechanism. Then what and how adaptions take place to the plants to react to the environmental alternations are the first questions we want to

address. By understanding such self-regulating mechanism, we might be able to find a way to artificially interference the gene expression to satisfy our need in the future. Based on these, Studies for DNA-methylation induced by elevated CO_2 with genome-wide single-base resolution is in great need.

In our experiments, it has been observed that under different concentrations of CO_2, the A.thaliana showed differences in leaf development and flowering time. It indicates that, in Arabidopsis, elevated CO_2could enhance photosynthetic rate and reduce the stomatal density and conductance. Therefore it would result in a higher amount of plant mass and seeds(Teng et al., 2009). Evidence has also shown that elevated CO_2can affect root hair development through auxin signaling (Niu et al., 2011).We believe that higher concentration of carbon dioxide will impact on the plants development through the reaction of epigenetic mechanisms.

We aim to answer the following questions in our study: What DNA methylation alternations would take place in plants in responding to higher carbon dioxide concentration? How are differentially methylated sites distributed along the genome? Do such epigenetic alternations impact any gene expression?

We study the model plant Arabidopsis thaliana that has five chromosomes, totally around 157 mega base pairs. In order to detect the methylation pattern and significantly differentially methylated sites, we compared the DNA methylation status of strains grew with two different concentration levels of carbon dioxide (430ppm vs 810ppm). Furthermore, we plan to search for the functional genes that are located adjacent to the clusters of methylation variations. Since higher temperature and humidity may increase the ratio of photorespiratory loss of carbon to

photosynthetic gain (Long 1991), to isolate the effect of increased CO_2, conditions such as temperature and humidity have been controlled.

3 Experiment design and material

Seeds of Arabidopsis thaliana (ecotype Col-0) were imbibed at 4 °C with distilled water for 4 days to facilitate uniform germination. The seeds were planted in 1m3 mesocosms in greenhouse with low (430±50 μmol mol−1 or ppm) and high (810±50 μmol mol−1 or ppm) CO_2 concentrations. After five self-seeding generations, the sixth-generation plants were grown for study. For the plants grown under 810 ppm CO_2, the first inflorescence stems grew on the 70th day after plant germination, and one week later for those under 430 ppm concentration. The onset of flowering in these plants occurred within 5–6 weeks. A sample of 100 uppermost fully expanded rosette leaves were randomly collected from plants at the 10 leaves stage. Samples from the 810 and 430 ppm CO_2 conditions were collected at the same time, about 1 or 2 weeks before first inflorescence stems were observed, respectively. Leaves were stored in liquid nitrogen at −80 °C for further analyses. The means based on all CO_2 concentrations recorded during plant growth indicate that the actual concentrations are indeed 430 and 810 ppm (May et al., 2013). Illumina gene sequence analyzer combined with bisulfite-treated conversion was used to obtain the single-end sequencing reads.

The reference sequence (version TAIR10) was obtained from the Arabidopsis Information Resource (TAIR), a database of genetic and molecular biology data for Arabidopsis thaliana. TAIR includes the complete genome sequence along with gene structure, gene product

information, gene expression, DNA and seed stocks, genome maps, genetic and physical markers, publications, and information about the Arabidopsis research community (Rhee et al., 2003).

For differentially expressed gene analysis in RNA-seq, total RNAs were extracted from frozen leaves. RNA-seq libraries were prepared using TruSeq RNA Sample Preparation v2 kit (Illumina Inc.). The sequencing was processed by Illumina Hi-seq2000 Sequencer. Sequencing data was deposit in Gene Expression Omnibus (GEO, GSE36934).

4 Method

The pipeline of our DNA methylation analysis is straightforward. The major steps are data processing, statistical significance test and Biological inference (Figure 8).

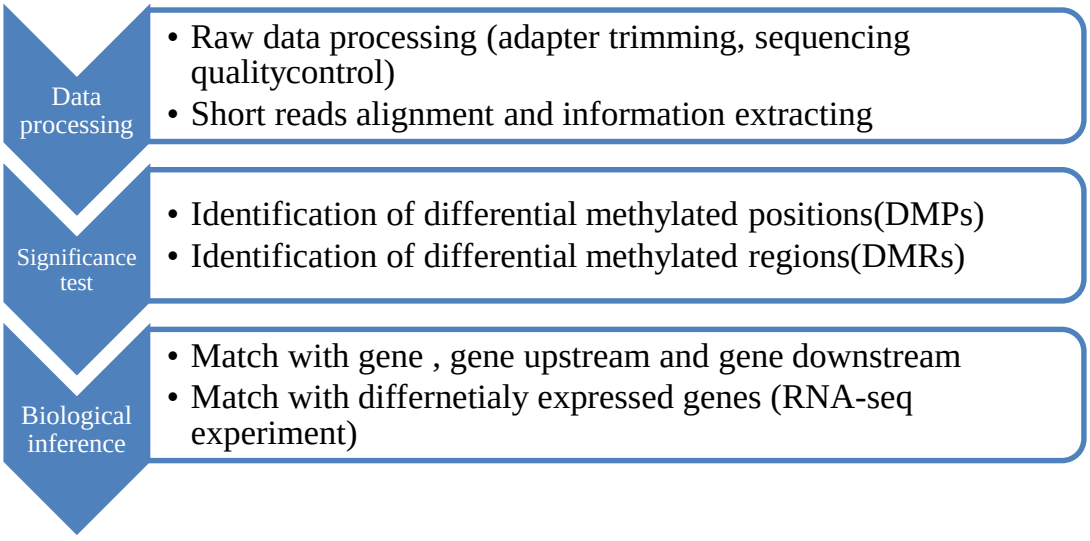


Figure 8 DNA methylation analysis pipeline. There are three major procedures in the analysis pipeline. In data processing step the raw sequencing data was preprocessed and mapped to the reference genome. Then the information for downstream analysis was extracted. The significance test was applied to the processed data to get the differentially methylated positions

(DMPs) and regions (DMRs). At last, the conclusion was drawn by locate the adjacent genes to DMRs.

4.1 Short read mapping and data pre-processing

Raw reads from Illumina genome analyzer were in fastq format, with sequencing ID, the read sequence and the quality scores for each nucleotide. After quality filtering and mapping to reference genome, the mapped data in BED format would contain the information of the mapped regions' coordinates on genome. The cytosine base calls were processed by Perl scripts to extract the "C" reads and the coverage information at each single cytosine position. Three types of cytosine contexts (CG, CHG and CHH) were defined in the base calls data.

A well-developed package FASTX_Toolkit (Gordon et al., 2010) was used to process the raw reads under quality control threshold. The reads with length less and equal to 10bp were filtered out. The adapters and low quality sequences were trimmed from both 5-prime end and 3-prime end. We used RMAPBS (Smith et al., 2008; Smith et al., 2009), which was built for bisulfite treated sequences alignment. By default, RMAPBS allows at most 4 mismatch for first 32bp of the reads. Some reads could be mapped to multiple regions on the genome, due to repeated sequences, or the short read length. In the following analysis, we only used the uniquely mapped reads.

At each cytosine position, we counted the number of methylated reads and unmethylated reads. Considering that the read coverage might impact the statistical power, we excluded positions with read depths less than 3.

4.2 Genome-wide multiple test

Pair-wise hypothesis tests were applied to identify significant differences between two samples. In genome wide studies, thousands of features, such as genes, specific regions or nucleotides, are involved in the significance test simultaneously. The false positives may occur under this circumstance.

False-positive occurs when features that show statistical significance (e.g. with p-value <=0.05) are truly from the null hypothesis. When we apply multiple tests independently, the incidence of false positives is proportional to the number of tests and the significance level. The traditional inference with individual p-value cutoff is no longer suitable. There are several approaches to control the portion of the false positives, such as Bonferroni, Bonferroni Step-down, Westfall and Young Permutation and Benjamini and Hochberg False Discovery Rate (FDR). Usually, the more stringent a method is, the less false positives will be called; however, the stringent method has the risk of inducing more false negatives (truly significances that are concluded as no significances).

In our study, two procedures need the adjustment for the multiple tests: defining the methylated cytosine and identifying the differentially methylated positions (DMPs). Defining the methylated sites was to further extract the cytosines we were interested in the following differentially analysis. Such that, we applied Bonferroni Step-down correction to retain a moderate amount of positives. While for identifying the DMPs, more significant features would benefit us to further study the differentially methylated regions. Therefore the Benjamini-Hochberg FDR correction was applied in this step. Details about the application are described in the following sections.

4.2.1 Define the methylated sites with Bonferroni Step-down correction

In this study, we wanted to test differences in the DNA methylation levels between two samples at every testable cytosine position. For a particular position, there may exist two types of differences: one strain was methylated while the other strain was not; the methylation level in one methylated strain was higher than the methylated strain.

Since both the bisulfite treated sequencing and the short reads alignment allow errors, some methylated reads may be due to un-conversion during the sequencing or mismatches in alignment. Therefore, when determining the cytosine is methylated or not, we have to tell if a methylated read is truly methylated first, which is determined by a Binomial probability.

In mammal study, as the methylation rarely happen on CHG and CHH sites, we can count the reads of cytosine as methylations. The genomewide false positive methylation rate is roughly calculated as the number of methylated reads at CHG and CHH divided by the total read counts that cover the cytosine. However, methylation could happen to all kinds of context in plants. Fortunately, in the chloroplast, DNA methylation would not occur. Therefore, the false positive methylation rate can be estimated by counting the methylated reads in sequences mapped to chloroplast genome. The error rate for each sample was calculated in this study.

At a given cytosine position, with the specific total number of covered reads and error rate, an probability of observing the methylated reads cover that position can be obtained by binomial distribution $\binom{n}{k} p^k (1-p)^{n-k}$, where n is the read depth, k is the methylated read count and p denotes the error rate. We treat this probability as the individual p-value: if this probability is less than a threshold we can call it a methylated cytosine.

Bonferroni Step-down correction is applied in this step, as we want an intermediate control for potentially methylated positions. Bonferroni Step-down correction is similar to the Bonferroni procedure. However, instead of multiplying the p-value with the total number of tests and teat each test with same weight, Step-down correction treat the p-value with different correction according to the rank of the p-value. The scheme of the Step-down correction is as below:

1) Sort the individual p-values in an ascending order.
2) Corrected p-value = original p-value *(n-rank+1), where n is the total number of tests.

Compare the corrected p-values with the significance level (eg. 0.05) , features with corrected p-values less than the threshold are considered to be significant. Both Bonferroni and Bonferroni Step-down correction aim to control the Family-wise error rate (FWER). FWER allow very few occurrences of false positives. For example, if FWER equals to 0.05, it is expected that 0.05 of features would be significant by chance.

Another potential algorithm of defining methylation cytosines was described in Lister,R,. et al (Lister et al., 2009). They adjusted the p-value according to observations at every position. To ensure the false discovery rate (FDR) of 0.05, the FDR controlled threshold for p-value at each position was calculated as $0.05*k/(n-k)$, where n is the read depth, k is the methylated read count. Using this algorithm, it is expected that out of all methylated cytosines called, no more than 5% would be due to the un-conversion and sequencing error. At each cytosine position, the observed binomial probability was calculated by a null distribution: $Binomial\ (n, error\ rate)$. Then a cytosine site with binomial probability less than FDR corrected P-value was called as methylated position.

We have applied both methods to our data and they showed no differences. So we retained the result from Bonferroni Step-down correction for following analysis.

4.2.2 Define the differentially methylated positions (DMP) with Benjamini-Hochberg FDR correction

To test the equality of methylation levels in two samples, the two-tailed Fisher's exact test was applied to 2×2 contingency tables derived at each cytosine position (Table 1). Let π_1 and π_2 denote the proportions of the methylated cytosines (mC) to the total reads (total read) covered at the tested position in two strains. The null hypothesis was $H_0: \pi_1 = \pi_2$.

	mC read	non-mC read	total read
Sample1	k1	n1-k1	n1
Sample2	k2	n2-k2	n2

Table 1. 2x2 table for differential test at each cytosine position. n1 and n2 denote the total coverage of the single cytosine position in sample1 and sample2; k1 and k2 indicate the methylated C read counts in two samples.

In testing the significance of association between methylation status and treatments, the Benjamini-Hochberg FDR correction was applied in this procedure.

1) At testable positions on genome (m positions in total), individual p-values ($P_1, P_2, \ldots, P_m$) were obtained by statistical test. The p-values then were sorted in ascending order ($P_{(1)}, P_{(2),} \ldots, P_{(m)}$).

2) For a given false discovery rate α, find the largest k such that $P_{(k)} \leq \frac{k}{m}\alpha$.

3) Reject the null hypothesis for the test 1 to k.

The retained positive results were then claimed to be genome-wide significant. Benjamini-Hochberg FDR is the least stringent method compared to the other three options. FDR allows a percentage of called significance to be false positives. If a FDR equals 0.05, we can expect 5% of significance features are identified by chance only.

A window sliding procedure was used to find the clusters of DMPs that identified by Fisher's exact test. The 50bp windows were initially obtained by satisfying the threshold that at least 3 DMPs in the 50bp windows. Then the windows were merged if the 50bp windows were overlapped. Regions with high frequency of DNA methylation variations may potentially impact their adjacent genes.

4.3 Biological inference

4.3.1 Differentially expressed gene analysis with RNA-seq data

In order to identify potential targets whose expression were influenced by DNA methylations. We matched the DNA methylation analysis with the RNA-seq analysis. We ran the

differentially expression analyses on three pairs of samples (430ppm vs 810 ppm CO_2 concentration), and a classic RNA-seq analysis pipeline (Figure 9) was applied to determine differentially expressed genes.

After adapters were trimmed, the paired-end raw RNA-seq reads were uniquely mapped to Tair10 reference genome using the TopHat package(Trapnell et al., 2009). Then, Cufflinks was used to assemble transcripts and estimated the relative abundances of possible isoforms. In essence, Cufflinks implements a constructive proof of Dilworth's Theorem by constructing a covering relation on the read alignments, and finding a minimum path cover on the directed acyclic graph for the relation. Cufflinks tries to find the correct and parsimonious set of transcripts by performing a minimum cost maximum matching. With normalized RNA-seq fragment counts, the Cufflinks measures the abundances in Fragments Per Kilobase of exon per Million fragments mapped (FPKM) (Mortazavi et al., 2008; Trapnell et al., 2009; Trapnell et al., 2012).

The assembly files are merged with reference transcriptome annotation into a unified annotation and fed into Cuffdiff for differentially expression tests. Cuffdiff calculates expression in two or more samples and tests the statistical significance of each observed change in expression.(Trapnell et al., 2010). It tests the observed log-fold-change in its expression against the null hypothesis of no change. Upper-quantile normalization was applied to two samples within each replicate to make two samples comparable in both Cufflinks and Cuffdiff option setting.

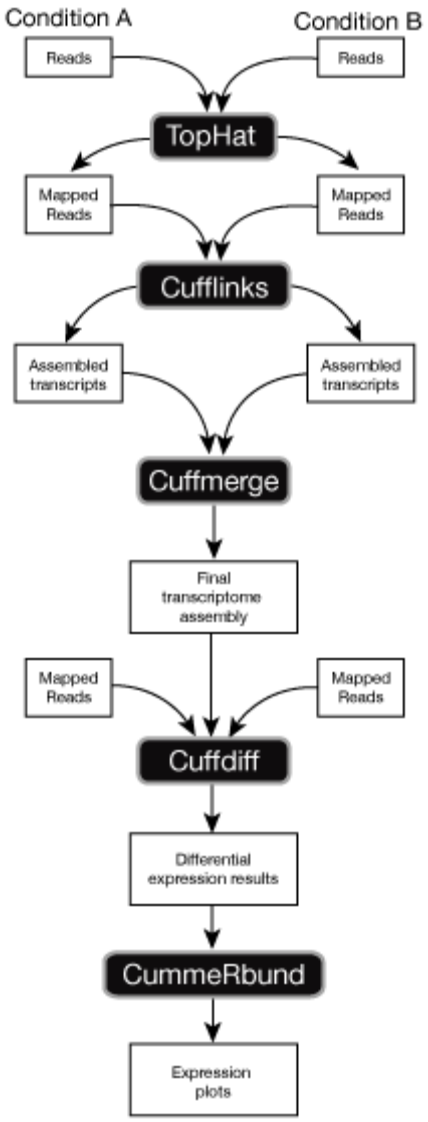


Figure 9. RNA-seq analysis pipeline (Galaxy tools). In a two conditions experiment, reads are first mapped to the genome with TopHat. These mapped reads are provided as input to Cufflinks, which produces one file of assembled transfrags for each sample. The Cuffmerge then merges the assembly files with the reference transcriptome annotation into a unified annotation for further analysis. This merged annotation is quantified in each condition by Cuffdiff, which produces expression data in a set of tabular files. These files can be indexed and visualized with CummeRbund to facilitate exploration of genes identified by Cuffdiff as differentially expressed, spliced, or transcriptionally regulated genes.

4.3.2 Map the DMRs with gene elements and differentially expressed genes

To explore the correlation between DNA methylation and gene expression, the distribution of DMRs along the gene body, five-prime UTR, three-prime UTR and promoters was studied. The DMR that has longer than 20% of its length overlapped with the element is considered as overlapped with the element. The .gff files with annotation information was obtained from TAIR10 (Rhee et al., 2003). Promoter regions that were defined as 1kb upstream of the start of five-prime UTR and overlapped with DMRs were also identified. There was no threshold for the overlapping length in match the DMRs with promoters.

We matched the gene list derived from above with the differentially expressed genes from RNA-seq analysis, and derived a group of overlapping genes. These genes are potentially impacted by the high DNA methylation level in reacting to the elevation of CO_2 concentration.

5 Results

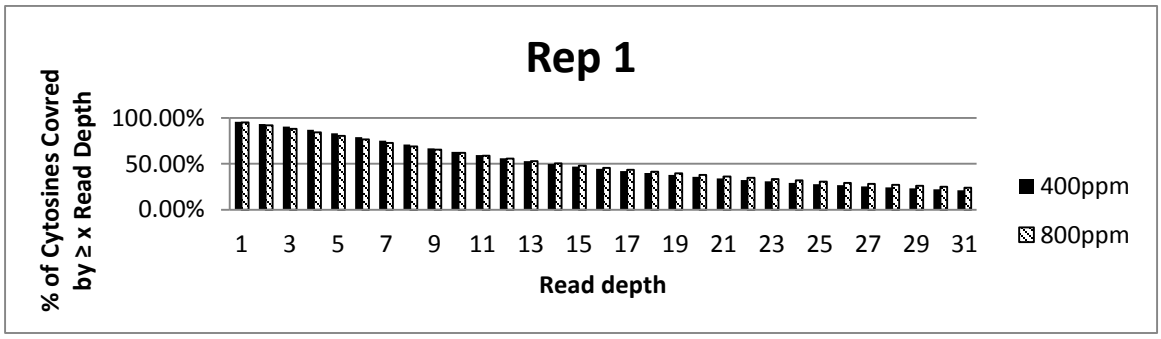
5.1 Bisulfite treated sequencing short reads alignment

We had ~150,000,000 short reads range from 4bp to 101bp long for each sample (Table 2). After quality control, adapter removal and uniquely mapping, around 70 million short reads were left for the downstream analysis. The mapped reads covered over 90% of the whole genome. The average read depth at each covered genome position was around 30 which indicate the sequencing result could satisfy the hypothesis testing (Table 2).

	Rep1		Rep2	
	400ppm	800ppm	400ppm	800ppm
Raw reads	182,456,168	150,138,348	152,093,025	129,331,514
Total reads after trimming	132,654,863	121,999,534	150,992,516	128,253,172
Uniquely mapped total reads	86,238,563	67,336,234	73,925,270	56,808,855
Genome average read depth (number of reads/base)	32.78	28.66	32.75	24.41

Table 2. Number of reads in data processing step and average genome coverage for four strains.

Over 90% of the cytosines on the reference genome were covered by sequencing data (Figure 10). In our hypothesis analysis for DMPs, we only kept the cytosines that were covered by at least three reads for the downstream analysis. As the result, ~85% of the genome cytosines were taken into analysis (Figure 10).



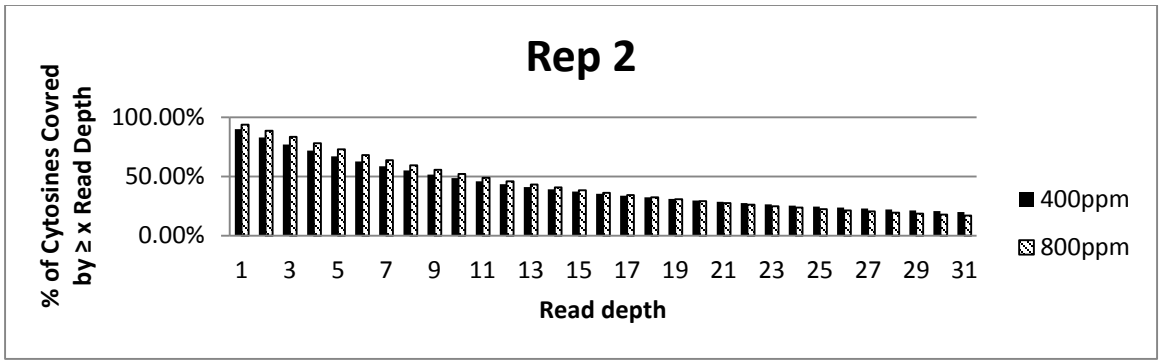


Figure 10. Distribution of genomic cytosines coverage. The figure was obtained by count the percentage of cytosines that were covered by at least one read to at least 30 reads. Four samples showed similar pattern of cytosine coverage. The percentages were averaged across five chromosomes in each sample.

5.2 Methylcytosine profiles

Our preliminary High Performance Liquid Chromatography (HPLC) experiments indicated that, genome-widely the cytosine would gain methylation with increased CO_2 concentration. Although we observed slightly decreased number of methylated cytosines from 400ppm strains to the 800ppm strains (Table 3), we expected large portion of cytosines with low methylation levels at 430ppm would turn to be highly methylated under 810ppm.

		Rep1 430ppm	Rep1 810ppm	Rep2 430ppm	Rep2 810ppm
Error rate		0.006417	0.008713	0.009667	0.010758
Binomial Tests (read depth ≥ 3)	Chr1	1,313,978	1,302,153	1,283,488	1,247,185
	Chr2	1,306,470	1,294,369	1,285,641	1,257,128
	Chr3	1,363,903	1,355,220	1,336,494	1,310,199
	Chr4	1,108,888	1,104,089	1,087,900	1,068,024
	Chr5	1,351,164	1,339,882	1,322,229	1,292,137
	Total	6,444,403	6,395,713	6,315,752	6,174,673

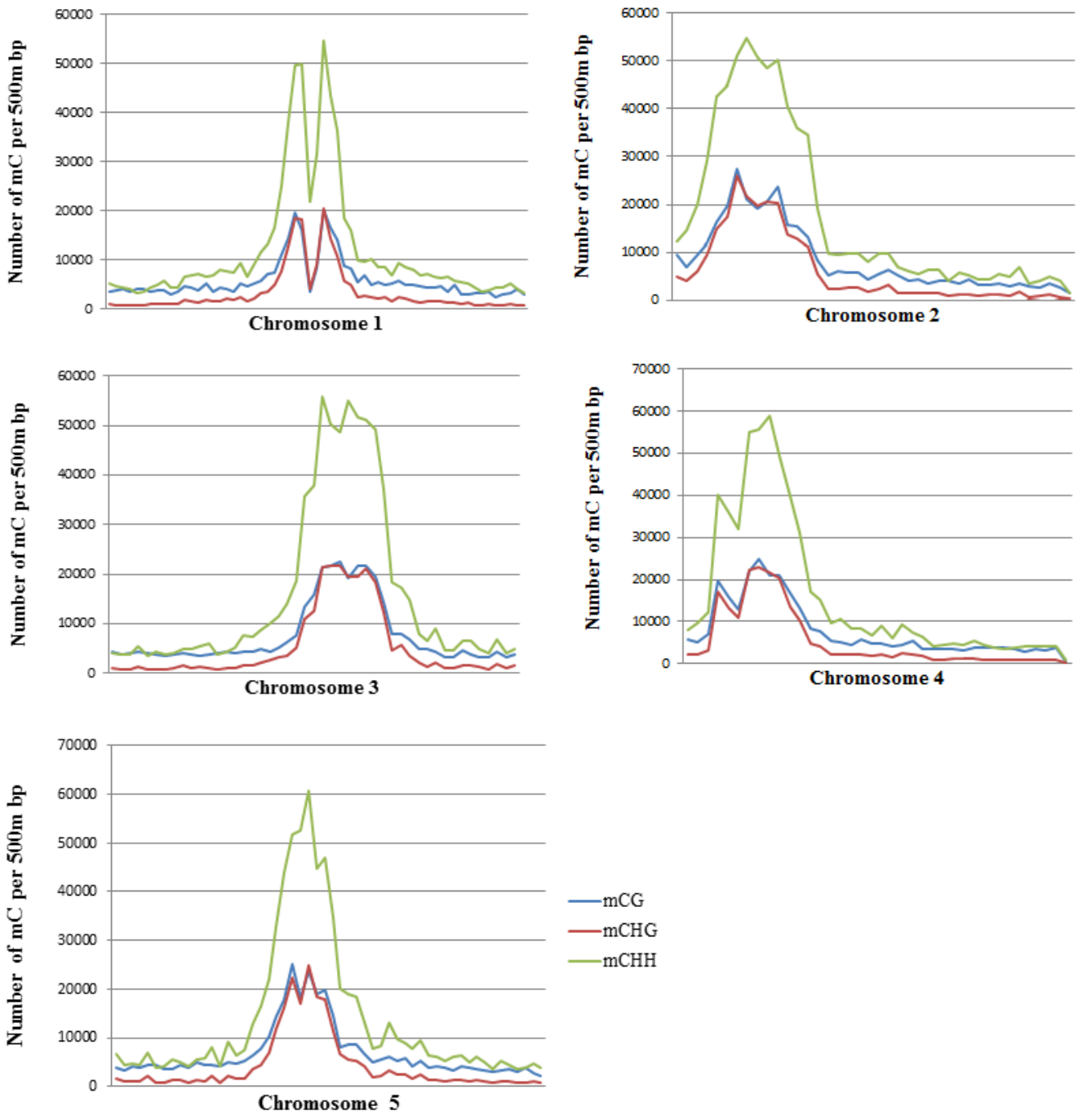
Table 3. Methycytosine profiles. The methylcytosine was determined by Binomial test with error rate specific for each sample.

Within each context, the portion of methylated CG sites was higher than the other two methylation forms (Table 4). Among the contexts of methylated cytosines, the mCHH sites were highly represented (Figure 11b), because the representation of CHH sites is much higher than the other two cytosine contexts in Arabidopsis genome. We did not observe big difference between the enrichment patterns of the methylated sites between two environmental conditions. The chromosome distribution of methylated cytosines shows that the methylation has much higher densities in the pericentromeric regions of chromosomes (Figure 11a). CHG methylation enriched in the pericentromeric regions, likely due to its preference for methylation of transposon-related sequences (Tompa et al., 2002; Kato et al., 2003). In contrast, CG and CHH methylation, although most occupied in the pericentromeric regions, showed relatively higher enrichment throughout the euchromatic chromosome arms (Figure 11a).

	Rep1 430ppm	Rep1 810ppm	Rep2 430ppm	Rep2 810ppm
CG	31.2%	31.5%	30.2%	30.4%
CHG	18.7%	18.9%	18.2%	18.3%
CHH	11.4%	11.2%	11.3%	10.8%

Table 4. Methylcytosine percentage within each cytosine context. Among three contexts, the proportion of methylated CG is the highest. Two strains at different $\mathbf{CO_2}$ concentration levels show similar patterns.

(a)


Number of mC per 500m bp
60000
50000
40000
30000
20000
10000
0
Chromosome 1
Chromosome 2
Chromosome 3
70000
Chromosome 4
Chromosome 5
mCG
mCHG
mCHH

(b)

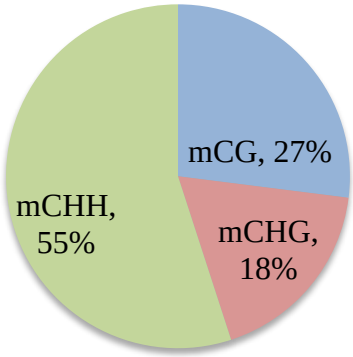


Figure 11. Methylated cytosines distributions of three cytosine contexts. (a) The distributions of three cytosine contexts have similar patterns along chromosome 1 to chromosome 5. The enrichment of methylcytosines elevated at the pericentromeric regions of chromosomes. The figure was obtained by split the chromosome into non-overlapped 500m bp windows and the average density of methylated cytosines was calculated within each window. (b) The methylated CHH was mostly representative among three contexts. The percentages were averaged across all samples.

5.3 Functional genes impacted by differentially methylated regions (DMRs)

We derived 921 DMRs from the differentially methylation analysis. Among these DMRs, 879 DMRs showed higher level of methylation in 810ppm strain. We found 45 genes that located overlapped with DMRs and 69 genes located within the 1kb downstream of the DMRs (Table S 1). Some of these genes were identified by the RNA-seq analysis as differentially expressed genes (FDR <=0.05) (Table 5). The genes overlapped with DMRs at the 5'UTR or 3'UTR did not show significance in differential expression. Three replicates did not show exactly same results but had common genes (eg. AT2G22300, AT2G03980). Theoretically, the DNA methylation level is positive related with gene expression at the genebody and negative related at

the promoter region. In our results, seems the association at the genebody shows consistently positive but the negative association at the promoter regions is not always true. This suggests that gaining methylation do not always repress the downstream gene expression. Some listed genes are found functional in the plant development.

RNA_seq replicate	Alignment	Gene locus	Sample	FPKM (430ppm)	FPKM (810ppm)	q_value
1	Genebody	AT2G01008	810ppm	2.22	9.22	2.71E-09
		AT3G41761	810ppm	0.00	71.45	9.36E-04
	Promoter	AT2G03980	810ppm	28.98	60.52	2.02E-04
		AT2G22300	810ppm	39.88	27.35	5.68E-03
2	Genebody	AT3G17185	430ppm	4.93	18.77	5.62E-03
	Promoter	AT2G03980	810ppm	19.27	55.86	3.03E-02
		AT2G22300	810ppm	32.24	14.28	5.00E-03
		AT4G08290	810ppm	22.66	7.28	8.11E-03
3	Genebody	AT5G24270	430ppm	2.35	5.41	3.43E-03
	Promoter	AT3G23790	810ppm	14.89	27.27	7.67E-03
		AT5G35732	810ppm	5.49	15.59	2.65E-03

Table 5. The genes that may impact by high frequency of DNA methylation variations. The overlap and promoter denote the gene is overlapped with DMRs and located at the 1kb downstream of the DMRs respectively. Most of the identified genes were found overlapped with the DMRs. Sample column represents that the methylation levels increased under the specific $\mathbf{CO_2}$ concentration. Fragments Per Kilobase of exon per Million fragments mapped (FPKM) values for transcripts and the q_value for differntial test values were generated by Galaxy tools described in the method session.

Elevated CO_2 may affect the lipid metabolic process

Metabolism is the set of life-sustaining chemical transformations within the cells of living organisms. These reactions allow organisms to grow and reproduce, maintain their structures, and respond to their environments. The gene AT2G03980 was a GDSL-like Lipase/Acylhydrolase superfamily protein which is related to the hydrolase activity. As it is involved in the lipid metabolic process, regulation of this type of genes played very important roles in the plant development, especially under certain environmental stresses. We found this particular gene showed around two-fold change in expression at 800ppm CO_2 concentration, compared to the 400ppm CO_2 concentration (Table 5). This change may be correlated with the cluster of increased methylation at the 1kb promoter region.

Besides, we observed a significant higher expression of AT3G23790 with increasing methylation level under elevated CO_2 (Table 5). The AT3G23790 is a gene that could produce acyl activating enzyme 16 (AAE16) which is also involved in the metabolic process (Koo et al., 2005). The Arabidopsis genome contains a superfamily (63 members) of genes encoding proteins annotated as acyl-activating enzymes. These enzymes catalyze the activation of many carboxylic acid substrates through the formation of thioester bonds. Along with another member AAE15(At4g14070), AAE16 was grouped into a long-chain acyl-CoA synthetase subfamily (LACS gene).

Elevated CO_2 may affect the defense responses

The differentially expressed gene AT2G22300 is one of the calmodulin-binding transcription activators (CAMTAs) family. The CAMTAs comprise a conserved family of

transcription factors in a wide range of multicellular eukaryotes, which possibly respond to calcium signaling by direct binding of calmodulin. Arabidopsis thaliana contains six CAMTA genes (At5g09410, At5g64220, At2g22300, AT1G67310, At4g16150 and At3g16940).

Loss-of-function mutations show enhanced resistance to fungal and bacterial pathogens suggesting that CAMTA functions to suppress defense responses (Rhee et al., 2003). Calmodulin (CaM) is a small (148-residue), highly conserved, ubiquitous, calcium binding protein. A number of CaM-binding proteins have been identified through classical methods, and many proteins have been predicted to bind CaMs based on their structural homology with known targets. CaM binds to proteins involved in the regulation of an array of cellular processes, including gene transcription, muscle contraction, cell survival, and neurotransmitter disease (Klee et al., 1980; Chin et al., 2000; Yamniuk et al., 2004). Earlier reports suggested that CaM activity could be regulated via methylation because the methylation state of CaM was observed to vary in a tissue-specific and developmentally specific pattern in Pisum sativum (pea) roots (Oh et al., 1990). A recent study further identified new methylation-dependent CaM binding proteins (Banerjee et al., 2013).

Elevated CO_2 may affect the auxin-regulated organ development

The gene AT3G17185 which overlapped with DMR showed higher expression with loss of methylation in responding to the higher level of CO_2 concentration. This gene encodes a trans-acting siRNA (tasi-RNA) that regulates the expression of auxin response factor genes (ARF2, ARF4, ETT) and it is one of 3 genomic loci that encode the TAS3 siRNA (Rhee et al., 2003). As it may be involved in the leaf development, it is definitely related to the adaption of the plants to the environmental impact like CO_2 concentration.

Elevated CO2 may affect the nutrition, K+/Na+ selectivity, and salt tolerance

Another gene, AT5G24270, is similar to calcineurin B. Lines carrying recessive mutations are hypersensitive to Na+ and Li+ stresses and is unable to grow in low K+. It encodes a calcium sensor that is essential for K+ nutrition, K+/Na+ selectivity, and salt tolerance. (Rhee et al., 2003). It has been demonstrated that chromatin modification is involved in the resistance responses of plants to salt stress in the same generation as the stress occurs. A study reported that failure of cytosine methylation at a putative small RNA target site of AtHKT1.1 promoter led to lower gene expression, resulting hypersensitivity to salt stress(Shkolnik‐Inbar et al., 2013). Furthermore, another study has been shown that salt stress induced demethylation of NtGPDL gene could lead to higher tolerance to stress (Pavet et al., 2006). In our study the demethyaltion at the AT5G24270 gene region was likely to impact the salt tolerance of Arabidopsis in higher concentration of CO_2.

6 Conclusion

In this study Arabidopsis were treated with two levels of CO_2 concentrations to determine how CO2 might affect their DNA methylation status, using the high-throughput sequencing data generated by the NGS techniques. Genome-wide multiple tests were applied to get the significant methylation variance. A group of functional genes were found to be potentially impacted by high frequency of gaining or losing methylation at the nearby cytosine sites.

Our study indicated that the epigenetic modification, especially the DNA methylation, do contribute to the plant's adaption to the higher CO_2 stress. It suggested that, DNA methylation

variances may impact the crucial pathways of metabolic process, defense responses, nutrition selectivity and salt tolerance. This was consistent to the observed phenotype differences in flowering, leave development and seed production under two different levels of CO_2.

Together with other existing studies of stress-induced epigenetic modifications, our study supports further exploration of how increasing greenhouse gas may influence plants. This is important in future agriculture research and will ultimately facilitate us to meet the challenges brought by the global climate change.

Chapter III: Genome-wide association study between gene regulation and histone modification

1 Previous studies to detect the correlation between epigenetic alternations and gene expression

Histone modifications add another layer of epigenetic regulatory option. As more and more evidence shows that histone modifications are directly related with human disease like cancer, understanding how histone modifications are related to gene regulation is very important and meaningful, and would facilitate us to explore effective treatment. With the remarkable progress in this area, we begin to scratch the "code" of the histones but many problems remain to be solved.

Studies may explore qualitative and quantitative correlations between histone modifications and gene expression. Qualitative studies mainly describe the distribution of histone modifications tags around the gene regions, but quantitative studies aim to build predictive models for gene expressions based on histone modifications, for example, from the ChIP-seq enrichments. We will briefly introduce both types of analyses below.

1.1 Qualitative correlation studies

The link between histone modifications and transcription has been intensively studied. Most of them were focused on locating the density or frequencies of histone tags with the

corresponding active or repressed genes (Schones et al., 2008; Maunakea et al., 2010). Association can then be defined if the densities or signal distributions of chromatin features coincided with the various expressions of adjacent transcripts. It has been shown that histone acetylation is clearly associated with transcriptional activation (Maunakea et al., 2010), and the reverse reaction catalyzed by histone deacetylases (HDACs) results in transcriptionally repression and lowers the transcription potential of the underlying DNA (Braunstein et al., 1993). Different from acetylation, histone methylation could be associated with transcriptional activation or repression depending on the specific targeted residue and degree of methylation (Jenuwein et al., 2001; Kouzarides 2007).

It was suggested that histone modifications may delineate functional features of genes, including their structure and intrinsic regulatory elements. Histone modifications across actively expressed genes can be classified into at least 4 categories. Based on their general patterns, Histone modifications may correspond to different functions in transcription (Figure 12a). Inactive genes may be related to two types of patterns of "silent" histone marks (Figure 12b). It was shown that, "active" histone modification marks, which are highly enriched within gene promoters, may be involved in transcription initiation, whereas those at intragenic regions may be involved in elongation, termination or pre-mRNA splicing (Barski et al., 2007; Wang et al., 2008; Kolasinska-Zwierz et al., 2009; Wang et al., 2009; Luco et al., 2010). Similarly, other histone modifications across silent genes adopt distinct patterns to possibly impair/prevent transcription initiation when enriched at promoters, or to disrupt elongation when enriched throughout gene bodies (Maunakea et al., 2010).

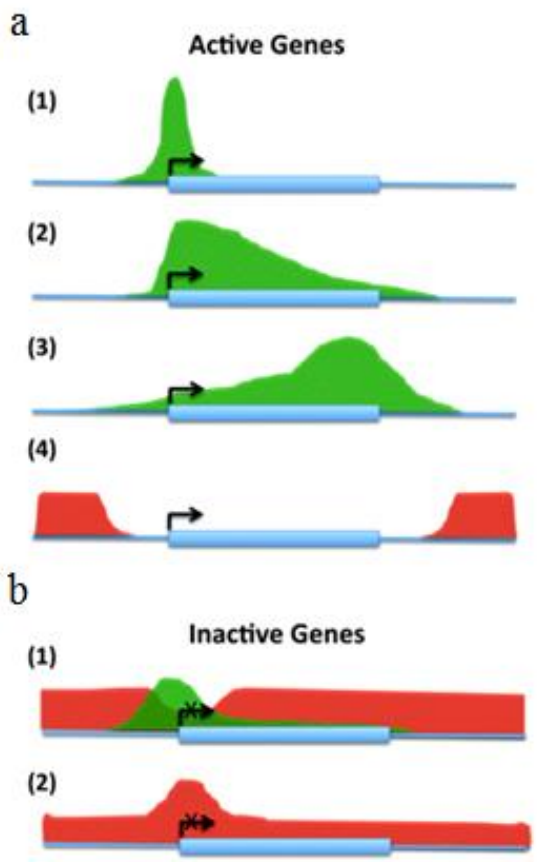


Figure 12. Distinct histone modification patterns delineate gene structure and associate with gene expression states. (Maunakea et al., 2010) Green color denotes the "active" histone modification marks, and red color denotes the "silent" histone modification marks. (a) Four categories of active histone modifications pattern relate to active genes. (1)Active mark highly enriches at gene promoters but depleted in gene bodies; (2) Active mark gradually decreases along gene bodies; (3) Active mark gradually increases throughout the gene bodies but peaks toward the 3'-end.(4) Both active mark and silent mark do not cover the gene region. (b) Two categories of silent histone modifications pattern relate to inactive genes. (1) Repression mark enriches over the gene even with an active mark surrounding the promoters; (2) Silent mark modestly enriches at promoters and declining signals throughout the gene region.

1.2 Quantitative correlation studies

Recent researches started to study the quantitative relationship between histone modifications and gene expression levels (Karlic et al., 2010; Xu et al., 2010; Cheng et al., 2011; Dong et al., 2012). In order to build quantitative models of gene expression on chromatin modifications, two important issues need to be considered. Firstly, how to define a proper variable from histone modification tags from ChIP-Seq enrichment data? Like other chromatin features, histone modification marks do not have consistent density distributions along the whole gene regions. How to quantify these patterns into variables for modeling is a difficult and yet critical question. Secondly, what model could be used to fit the data?

One study illustrated a quantitative model for predicting gene expression by histone modification using linear regression model (Karlic et al., 2010). The independent variables were log transformed sums of histone tags in 4001 base pairs surrounding the transcription start sites (TSS). Linear regression models were fitted with all possible combination of predictors (Figure 13). It was suggested that three histone modifications at the promoter are enough to faithfully model the expression of the associated gene. The Pearson's correlation coefficient r (Pearson's r) between predicted and observed expression level was computed to evaluate the model performance. The full model with all histone modifications yielded r=0.77. It reported that, combinations of only two (r_{max} = 0.74, H3K27ac + H4K20me1) to three modifications (r_{max} = 0.75, H3K27ac + H3K4me1 + H4K20me1) could account for 95% of the prediction accuracy of full model. According to the highest scoring models, some combinations of specific histone modifications were found to be always present in the model.

One –modification model (41 models)

$$f(X_i) = a + b_i X_i$$

Two-modifications model (820 models)

$$f(X_i, X_j) = a + b_i X_i + b_j X_j$$

Three-modifications model (10660 models)

$$f(X_i, X_j, X_k) = a + b_i X_i + b_j X_j + b_k X_k$$

Full model (1 model)

$$f(X_1, \ldots, X_{41}) = a + b_1 X_1 + b_2 X_2 + \cdots + b_{41} X_{41}$$

Figure 13. Linear model frameworks of gene expression on total tags of histone modification in a 4001bp window around TTS. (Karlic, Chung et al. 2010). X_i is log transformed total histone modification tags. All possible models from one predictor models to full model were considered. The Pearson's *r* was calculated for each model to access the prediction accuracy.

The ChIP-seq enrichment data show that histone modifications dynamically vary along the gene regions. Some studies pointed out that summarizing all histone modification tags in a specific gene region will underestimate the association with the gene regulation. To address this issue, the binning methods were developed in more recent studies (Cheng et al., 2011; Dong et al., 2012).

1.2.1 Initial binning method

Cheng et al. (2011) developed the initial binning method to take the dynamic density distribution of chromatin features into consideration. They split the ±4kb window around the transcription start sites (TSS) and the transcription termination sites (TTS) into 100bp long bins (Figure 14), so there were a total of 160 bins for each gene. The mean density of histone modification tags was calculated for each bin.

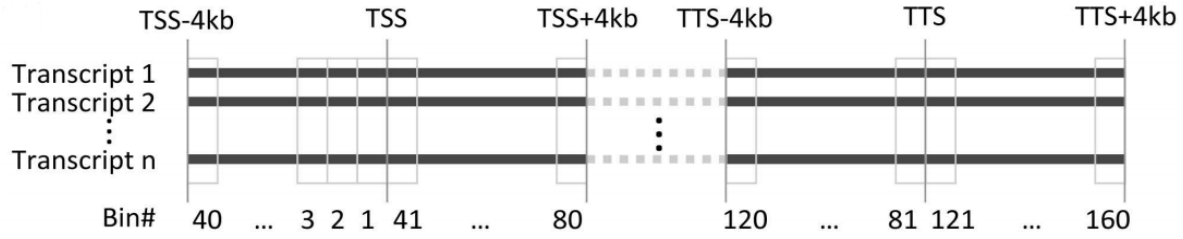


Figure 14. Initial binning method for quantifying the ChIP-Seq tags of histone modifications. ±4kb region around TSS and TTS were split into 100-bp bins (Cheng et al., 2011). Mean density was calculated in each bin to predict the gene expression.

Gene expressions were categorized into high and low expression groups based on RNA-seq data. Average signal of twelve histone modifications and other chromatin variants (such as Poly II) were calculated and used as predictors in the Support Vector Machine (SVM) to classify genes into the two expression groups. The binary classification was fitted in each of the 160 bins. Models were evaluated in terms of prediction performance by the area under the curves (AUC) in their receiver operating characteristic (ROC) curves. It has been reported that all bins were useful to classify gene expression but they were not equally informative. Among the 160 bins, those close to the TSS or TTS were more informative than those far away (Figure 15). It was

found that the regions around TSS (-300 to 500bp) or upstream of TTS (-200 to 0bp) had the highest classification accuracy.

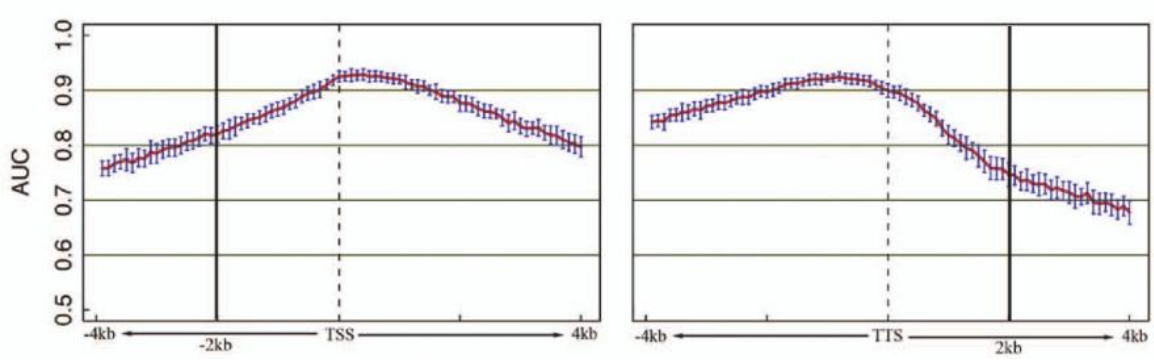


Figure 15. The prediction accuracy (AUC values) of SVM classification models for all the 160 bins along TSS and TTS regions (Cheng et al., 2011). It indicated that the bins close to the TSS and TTS are more predictive.

By investigating the individual or a subset of features, it was claimed that the model with 9 features could achieve almost the same prediction accuracy as the one with all 16 features. Furthermore, the authors trained the SVM model on subset features of methylation on histone H3 (K4, K9, K36 and K79). Among all individual predictors, methylation on H3K79 was found to be the most informative.

Linear regression models were also fitted for each bin to predict the continuous gene expression levels directly in a C.elegans dataset. For the most predictive Bin #1 (the closest bin at the TSS upstream), the Pearson's r reached 0.75. The method has also been applied to other organisms, such as yeast, fruit fly, mouse and human. Specifically, the Pearson's r reached 0.73 for applications to the human K562 cell line dataset.

In order to identify interaction effects, the authors modeled the expression level with a linear combination of all individual features and their two-way products. Among all 66 possible

interactions, some interactions were statistically significant. However by comparing the interaction model with the model with only main effects, the interaction terms did not substantially contribute to the prediction accuracy.

1.2.2 Best-bin method

To get better estimation with the representative signals for each chromatin features, Greven et al. improved the binning method by searching for the "bestbin" for each histone modification (Dong et al., 2012). The binning method in this study was similar with the initial binning method, except that only one bin was chosen as the representative signal for each feature. This study narrowed the region of study to $\pm$2kb window around the TSS and TTS. The specific gene region was divided into 81bins (Figure 16). The "bestbin" was defined as the bin with the strongest correlation with the expression level.

Figure 16. The Best-bin method split the genetic region into 81 bins(i.e. 40 bins for [-2k, +2k] of TSS, 40 bins for [-2k, +2k] of TTS, and 1 bin for the rest of gene body) (Dong et al., 2012). The mean density was calculated in each bin. The best bin was pick out by comparing the correlation coefficient between the histone modification signal and the gene expression across all 81 bins.

This paper proposed a two-step model. They first used classification models to identify the expressed genes (genes with non-zero expression level) versus un-expressed genes (genes with zero expression level). Then they fitted a simple linear regression model on the expressed

genes stratified by classification model to predict the continuous expression levels. Although it was claimed that the two-step model improved the prediction performance, there were two major drawbacks. Firstly, there were 40% of the expression values equal to zero, which were excluded in the linear model step. That is, the linear model was built only on the non-zero expression values. In this case, the prediction error tended to be under-estimated. Moreover, as mentioned in the study, the important features for classification model may differ from the linear regression model, so it was difficult to draw a clear biological interpretation.

The author also considered about the nonlinear effect in prediction. It was claimed that with the single feature effect, simple linear regression could obtain similar prediction performance as the other non-linear models, such like MARS (multivariate adaptive regression splines). However, the interaction effects were not considered in this study.

The generalized applications in other cell lines were compared. They applied the two-step procedure on different cell lines and three expression profile platforms (CAGE, RNA-PET and RNA-seq). Among different expression profile platforms, the model involved the CAGE TSS-based expression levels were most predictive with the chromatin features. There were 78 expression experiments in total. The median of 78 Pearson's r in two-step model was 0.83. With the two-step model, the largest Pearson's r value (0.895) was using PolyA+ cytosolic CAGE RNA expression profiles of K562 cells. With the same dataset, the simple linear model got r = 0.871.

In order to evaluate the contribution of different types of chromatin features, they applied two-step modeling to several subsets of chromatin features. The chromatin features were grouped in to following categories based on their known functions (Raisner et al., 2005; Barski

et al., 2007; Benevolenskaya 2007; Koch et al., 2007; Steger et al., 2008; Kolasinska-Zwierz et al., 2009): promoter marks (H3K4me2, H3K4me3,H2A.Z, H3K9ac and H3K27ac), structural marks (H3K36me3 and H3K79me2), repressive marks (H3K27me3 and H3K9me3) and distal/other marks (H3K4me1, H4K20me1 and H3K9m). The prediction accuracy was determined on each predictor subset and combination of predictor subsets. They concluded that for CAGE TSS-based gene expression, promoter marks and promoter combined with other categories gave similarly high Pearson's *r* values. On the other hand, non-promoter marks had relatively low prediction accuracy especially with the repressive marks only. They also compared the "all histone modifications" model with and without Dnase I hypersensitivity. Although for CAGE expression with the Dnase I the prediction was slightly more accurate, they noticed that the Dnase I itself did not effect a lot in gene expression.

Besides the binning method, Xu Hoang and colleagues (Xu et al., 2010; Hoang et al., 2011) proposed another enrichment data quantification which weighted the chromatin features according to the genome coordinates and gene length. They claimed that by assigning different weight to the modification enrichment levels, the model could adopt the spatial deposition patterns. Therefore, the rescaled predictors could improve the prediction accuracy.

Both binning method and refinement of enrichment estimation have pointed out the importance of the genome coordinate dependent association between modification enrichment predictors and the gene expressions. However, these scenarios focused on the data quantifications. Although the binning methods achieved high prediction accuracy for gene expression, by taking total or average histone modification tags in specific window or one bin as the predictor still lost the information from other positions that may impact gene expressions

cooperatively. Additionally, either modeling in every bin (initial binning method) or a representative bin (Bestbin method) brought difficulty in interpretation.

Previous methods indicated that the spatial correlation patterns may potentially exist between histone modifications and gene expression levels. However, the synergic effect, which makes up the "histone code", was largely unexplored. The mechanism of interaction among the histone modifications and other chromatin variants remains unclear.

2 Main goal and significance of the study

The quantitative modeling revealed the association between modification and gene expressions. It opened a door for exploring the epigenetic gene regulation. Due to the complexity of dynamic densities distribution of modification marks, models based on summation or average of tags along gene regions may be biased. Therefore, binning method is a good way to capture the varying patterns of the enrichment. To well utilize the spatial information which are the genomic loci the bins located is a necessary extension.

In our association study, we want to address the following questions: What is the quantitative relationship between the dynamic histone modification patterns and the gene expression? Do interactions among different chromatin features contribute to the gene expression?

Many previous studies showed that the enrichments of the histone modifications were related to the gene expression. We hypothesize that some positions are more informative to predict the gene expression than others. Based on the histone code hypothesis, the combination of histone modifications happens at the specific gene regions would contribute cooperatively to

the gene expression. Therefore, we believe that different genome locus also have systematic effects in the epigenetic gene regulation.

The existing methods seem to have already achieved a good performance based on the accuracy. However, they usually provide poor model interpretation. If we are interested in understanding the underlying mechanisms about how data were generated, new method with better interpretability has to be developed.

Our goal is to explore the correlation between histone modification and gene expression in a comprehensive and spatial way. The spatial component effect will be considered in both the individual and interactive histone modifications. A model that can adequately describe the quantitative relationship would benefit the further epigenetic study, and provide crucial evidences for the epigenetic regulatory system to the gene expression.

3 Materials and method

3.1 Data description

Datasets used in this study were referred to the Bestbin study (Dong et al., 2012). The ChIP-Seq data for fourteen chromatin features were available from the Gene Expression Omnibus (GEO; accession number GSE29611); DNsase I hypersensitivity data was accessible via GEO accession number GSE32970. Gene expression raw data profiled by CAGE can be downloaded with GEO accession number GSE34448.

3.1.1 Gene expression data

The cell line or cell culture refers to different culture conditions. In our study, data were generated from the K562 cells. Gene expressions protocols from different cell compartments (whole cell, cytosol and nucleus) and RNA extractions (Long PolyA (PolyA+) and Long Non PolyA (PolyA-) were compared in Greven et al. study. Overall, PolyA+ RNA was more predictive than PolyA- RNA.

The cap analysis gene expression (CAGE) is a promoter-based expression profiling technique. It was designed to locate the exact transcription start sites in the genome. Thus it benefited researchers who aimed to investigate the promoter activities. As discussed in this Greven et al. study, the choice of gene expression profile platform did matter in the gene regulatory study. The promoter marks were more predictive to CAGE-based expression levels while the structural marks were more predictive for RNA-Seq expression

The RNA gene expression values of a given TSS is defined as the sum of the CAGE tags whose 5' end falls within the 101bp window centers on the TSS. The response values were $log2$ transformed. To avoid $\log_2(0)$, a small number 0.03 was added to the expression followed the Greven et al. (Figure 17).

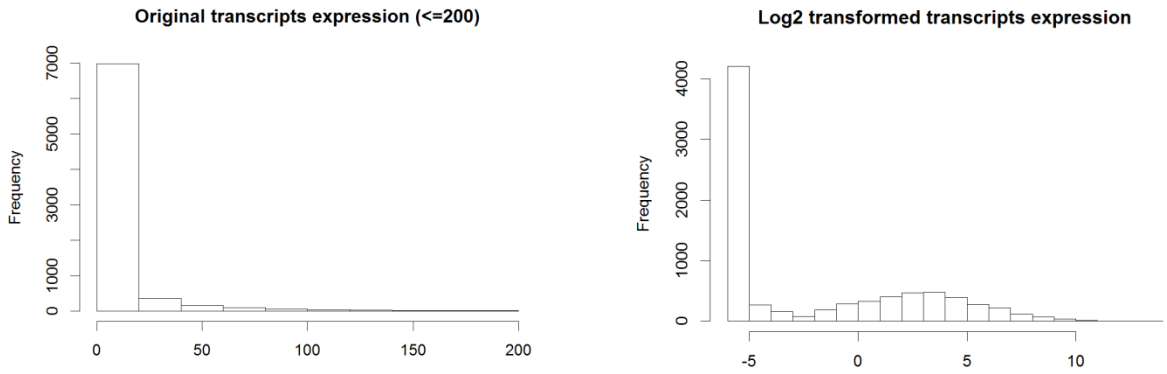


Figure 17. Histogram of gene expression values. The logarithm transformation of transcripts expression was applied to reduce the skewness.

3.1.2 Chromatin features data

Among all 14 chromatin features, 11 of them were histone modifications (H3k27ac, H3k27me3, H3k36me3, H3k4me1, H3k4me2, H3k4me3, H3k79me2, H3k9ac, H3k9me1, H3k9me3, H4k20me1). The rest two features were histone variant H2az and the Dnase I hypersensitivity. A control ChIP-Seq data corresponding to this cell line was also included as an individual feature. The number of accumulated mapped reads varied position by position and formed the dynamic patterns of the tag intensities.

Our analysis focused on the TSS sites of genes as the alternations in promoter regions were more likely to correlate with the gene regulations. Following the binning method, for each gene, the genomic regions that are $\pm$2k bp from the TSS was split into small bins of 100bp (40 bins in total). The region from the the 2kb downstream of the TSS to the 2kb upstream of the

TTS is treated as the 41st bin (Figure 18). Note that, the lengths of bin#41 may vary because of the different lengths of transcripts.

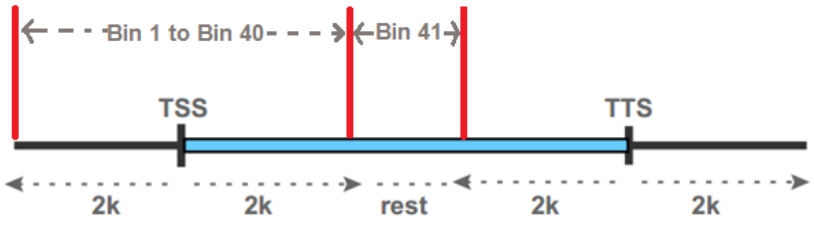


Figure 18. Quantification of the chromatin features. The 2kb region around TSS was split to 40 100bp bins. The Bin#41 is included as the gene body bin. Mean density of the ChIP-Seq tags was calculated in each bin to construct the predictor.

The number of ChIP-seq reads that covered the bin was counted and mean density was calculated in each bin with the bigWigSummary command-line utility(Kent et al., 2010). For the ith gene, jth histone chromatin and bin p, the original average density in that bin is denoted by $X^*j_p.i$(**Error! Reference source not found.**). As we have 41 bins in total, for the ith gene and jth histone modification, we had a vector $(X^*j_1.i \ \ldots \ X^*j_{41}.i)$ to represent the dynamic pattern of tags densities (Table 6).

The chromatin feature tags were also $log2$ transformed. A pseudocount was added to the original counts to avoid $\log_2(0)$. Following the data processing of Greven et al. study, one third of the dataset was taken to an optimization procedure in order to determine the pseudocount. Searching from 0 to 20% of the maximal value of X^*j_b in the b^{th} bin , the optimized pseudocount a_{j_b} was determined by a maximal correlation between $\log_2(X^*j_b.i + a_{j\_b})$ and expressions values.

After optimizing the chromatin features, we had 13,731 transcripts for modeling. The predictors were 14 chromatin features (CF), each having 41 bins for each transcript. The processed data has following structure (Table 6)

	Expression	CF1	CF2	…	CF14
Gene 1	Y_1	$X1_1.1, X1_2.1$, … , $X1_{41}.1$	$X2_1.1, X2_2.1$, … , $X2_{41}.1$	…	$X14_1.1, X14_2.1$, … , $X14_{41}.1$
Gene 2	Y_2	$X1_1.2, X1_2.2$, … , $X1_{41}.2$	$X2_1.2, X2_2.2$, … , $X2_{41}.2$	…	$X14_1.2, X14_2.2$, … , $X14_{41}.2$
…	…	… … …	… … …	…	… … …
Gene n	Y_n	$X1_1.n, X1_2.n$, … , $X1_{41}.n$	$X2_1.n, X2_2.n$, … , $X2_{41}.n$	…	$X14_1.n, X14_2.n$, … , $X14_{41}.n$

Table 6. Finalized data structure. The expression Y is $log2$ transformed expression values. $Xj_p.i$ denotes the $log2$ mean density of j^{th}chromatin feature at p^{th}bin in i^{th} transcript region. There are 41 observations for each of 14 chromatin features responding to each transcript.

3.2 Multivariate adaptive regression splines (MARS) and Step-wise multivariate adaptive regression splines (SMARS)

Multivariate adaptive regression splines (MARS) is a nonparametric nonlinear algorithm developed from adaptive computation strategies. These strategies are mainly used to approximate general functions in high dimensions. With the development of statistical methodology in this area, two major adaptive algorithms have been intensively studied. One is projection pursuit (Friedman et al., 1981; Friedman et al., 1983), and the other one is recursive partitioning (Breiman et al., 1984).

Projection pursuit is an approximation of M additive functions of linear combinations of variables

$$\hat{f}(x) = \sum_{m=1}^{M} f_m(\sum_{i=1}^{n} a_{im} x_i) .$$

The coefficients are estimated by joint optimization to reduce the square error loss. Projection pursuit regression could be viewed as a low dimensional expansion, which is adjusted to best fit the data. It was shown that, even with small size of M it could achieve enough approximation to many types of functions (Donoho et al., 1989). However, there exist some simple functions that need a large M for good approximation, and it also bring difficulties for interpretation. Moreover, in this case it is computationally intensive.

The recursive partitioning regression model is generally viewed as a geometrical procedure. It's also been viewed as a stepwise regression procedure. The approximation takes the form

$$\hat{f}(x) = \sum_{m=1}^{M} a_m B_m(x) .$$

The basis functions B_m take the form

$$B_m(x) = I[x \in R_m],$$

which is an indication function having value one if point x belong to the disjoint subregion R_m and zero otherwise. Given one observation x, only one function among all M basis functions would yield a nonzero value. The coefficients $\{a_m\}_1^M$ are jointly adjusted to get the best fit to the observed data.

The goal of recursive partitioning is to estimate a good set of subregions and parameters associated with the separate functions in each subregion. Starting with the entire domain D, the partitioning is accomplished by recursive splitting of previous subregions. The variables that locally have more influence on the response are more likely to be chosen as the optimal knot in splitting procedures. Globally, the response depends on a large number of variables but in each split subregion, it may only strongly depend on a subset of variables. Moreover, these subsets may vary a lot across the subregions. However even this is true, the recursive partitioning does not incorporate the subset selection feature in the algorithm. This limits the power and the interpretability of models. The other limitation is caused by the discontinuous functions at the subregion boundaries. When the true underlying function is continuous, this approximation would be bias. This problem is caused by using step function $I[x \in R_m]$.

In addition to that, it is not represented in the model whether the underlying function is a simple linear or additive model. It also has difficulty when interaction effects exist within a small fraction of variables. The recursive algorithm is to delete the existing basis function and replace

it with the product of basis functions. By doing so, the average interaction order increases after every loop of proceeding. Thus the model could not achieve good approximation when the function is composed of low order interactions. The additive model is a common case.

The feature selection of recursive partitioning is to remove the basis functions that do not contribute to the accuracy of approximation. The traditional way of deleting one basis at a time could not be used in the case of recursive partitioning. The corresponding regions are disjoint and removing a basis function will produce a "hole" in the predictor space. To trim the over-fitted model to a proper size, the backward strategy of recursive partitioning is designed to delete regions in adjacent pairs by merging them into a single region. One optimal algorithm is complexity tree pruning (Breiman et al., 1984).

3.2.1 Multivariate adaptive splines (MARS)

The MARS algorithm aims to adopt the adaptability of recursive partitioning while modifying it to overcome its two major limitations: inability to capture simple associations and discontinuity (Friedman 1991). The modifications include that:

1) Replace the step basis function by a truncated spline basis function $[\pm(x - t)]_+$

2) Not remove the existing "parent" basis function when new splits are included. Therefore, all basis functions could be involved for further splitting. Then the procedure is able to produce a model with high- or low-order of interactions. A simple function like linear or additive could be properly approximated.

MARS can be viewed as a generalization of stepwise linear regression or a modification of the CART method. The recursive partitioning regression procedure is well suited for high-dimensional problems. In addition of that, it is flexible enough to model non-linearity and variable interactions yet keep the interpretability. These advantages align well with the goals of our study.

The idea of MARS is to build an expansion form in a set of basis functions to cast the approximation. A good set of basis functions based on the data is derived from minimizing the lack of fit function (LOF). The piecewise basis functions are in form $(x-t)_+$ and $(t-x)_+$, where

$$(x-t)_+ = \begin{cases} x-t, & if\ x > t \\ 0, & otherwise \end{cases} \text{ and } (t-x)_+ = \begin{cases} t-x, & if\ x < t \\ 0, & otherwise \end{cases}.$$

Suppose we have features Xj's, $j = 1,2,\dots,q$. The knot value t is any possible value in observed values of feature Xj. Therefore, the collection of basis functions $C$is

$$C = \{(Xj-t)_+, (t-Xj)_+\}\ , t \in \{xj.1, xj.2, \dots, xj.n\}\ , j = 1,2,\dots,q$$

Note that although each pair of basis functions seems only to be based on one covariate Xj, the choice of basis function is considered over thte entire input space $\mathbb{R}^q$. The model then has the form

$$f(X) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X)\ ,$$

where each $h_m(X)$ is a function in collection C, or a product of two or more such functions.

Given a choice of basis function h_m , the coefficients β_m are estimated by minimizing the residual sum of squares. To choose an optimized set of basis functions is the core model

building procedure of MARS. Start with a constant function $h_0(X) = 1$, a new pair of basis function is selected from collection of candidate functions C at each step.

Using example given by (Hastie et al., 2009), suppose that after the first step we have model

$$\hat{f}(X) = \hat{\beta}_0 + \hat{\beta}_1(X_2 - x2.7)_+ + \hat{\beta}_2(x2.7 - X_2)_+ ,$$

when we add the new basis function pair, we have three options

$$\hat{f}(X) = \hat{\beta}_0 + \hat{\beta}_1(X_2 - x2.7)_+ + \hat{\beta}_2(x2.7 - X_2)_+ + \hat{\beta}_3\left(X_j - t\right)_+ + \hat{\beta}_4\left(t - X_j\right)_+ ,$$

$$\hat{f}(X) = \hat{\beta}_0 + \hat{\beta}_1(X_2 - x2.7)_+ + \hat{\beta}_2(x2.7 - X_2)_+ + \hat{\beta}_3(X_2 - x2.7)_+\left(X_j - t\right)_+ + \hat{\beta}_4(X_2 - x2.7)_+\left(t - X_j\right)_+ , \text{ or}$$

$$\hat{f}(X) = \hat{\beta}_0 + \hat{\beta}_1(X_2 - x2.7)_+ + \hat{\beta}_2(x2.7 - X_2)_+ + \hat{\beta}_3(x2.7 - X_2)_+ \cdot \left(X_j - t\right)_+ + \hat{\beta}_4(x2.7 - X_2)_+ \cdot \left(t - X_j\right)_+ .$$

Generally, with a model set M we already have from previous steps, the new pair of basis functions to be added in the next stage is in the form

$$\hat{\beta}_{M+1}h_l(X) \cdot \left(X_j - t\right)_+ + \hat{\beta}_{M+2}h_l(X) \cdot \left(t - X_j\right)_+, h_l \in M .$$

Here, $\hat{\beta}_{M+1}$ and $\hat{\beta}_{M+2}$ are the least-squares estimators. The h_l could be 1 and the model is still additive. The basis function pair that gives the smallest SSE is chosen. In practice, with large features dimension, we can predefine the model size before fitting process for computational efficiency. This process is continued until the model set M gets the preset maximum number of terms.

The LOF is involved in both forward stepwise and backward stepwise procedures. In measuring the distances between the predictions and observations, the squared-error loss was used to obtain the data based estimations. As MARS procedure highly depends on the data, the flexibility of model could usually achieve low bias of estimates. However, with additional parameters, the variance of model is still very high. In minimizing the loss function

$$\Delta\left[\hat{f}(X), f(X)\right] = [\hat{f}(X) - f(X)]^2 ,$$

the model selection criterion is modified by adding penalty to the number of terms (basis functions). As the subregions corresponding to basis function are overlapped, the traditional way of removing one at a time strategy could be used. The term whose removal improves the fit the most or degrades it the least will be deleted. The optimal number of terms λ in the model could be chosen by minimizing generalized cross-validation in favor of computational saving (Friedman 1991).The number of terms in model therefore is determined by this generalized cross-validation criterion proposed by Craven and Wahba (Craven et al., 1978),

$$GCV(\lambda) = \frac{1}{N}\sum_{i=1}^{N}\frac{\left[y_i - \hat{f}_\lambda(x_i)\right]^2}{\left[1 - \frac{C(\lambda)}{N}\right]^2}.$$

Notice that this criterion is the average-squared residual of the fit to the data with the penalty to trade-off the increased variance due to the increasing model complexity.

The effective value $C(\lambda)$ counts for both the number of terms in the models and the number of parameters in choosing the optimal knots locations. If there are r basis functions and K knots are selected in the forward stepwise process, the value $(\lambda) = r + cK$. The value of c could be estimated with bootstrap or cross-validation. Over all situations of studies, the best

value of c is between 2 to 4. Particularly, in the R package {earth}, if the model was preset as additive model then c equals to 2 otherwise it equals to 3.

Due to the splitting by knots, the piecewise linear basis functions are able to operate locally. The regression surface is built up parsimoniously only where they are needed. This is particularly important when a high-dimensional data is involved. The forward step-wise modeling strategy is also a key advantage of MARS. Higher-order multiway products are always based on the existence of lower-order components. When doing the MARS in practice, we need to set the limit of order of the interaction. With an order equals 1, a model is just an additive model. With the backward feature selection step, MARS could automatically yield a good bias and variance trade-off.

In summary, there are advantages of MARS algorithm that can fit our goal of study. Firstly, MARS is much flexible than linear regression with less assumption for the distribution of data. Secondly, the piecewise linear basis function can operate the interaction term locally without introduce too may parameters in the model fitting. Thirdly, with the backward feature selection step, it can yield a good balance between model complexity and prediction accuracy.

3.2.2 Stepwise MARS (SMARS)

As described in the goal of our study, we aim to identify the chromatin modifications' effect from different genome locus. Moreover, we also want to explore the potential existence of the interactions between the modifications. In order to detect the combinational and spatial effect from chromatin features, we proposed the stepwise multivariate adaptive regression splines to

incorporate the potential interactions among different chromatin features and among different genomic loci. One can include the whole data set with all the features at all bins at one time then use model selection to find the final model. However, this would be computationally ineffective. More importantly, the model may have histone modifications from too many bins and it will bring difficulties for interpretation.

The Stepwise MARS (SMARS) is a data driven idea. By adding features at one bin as a new set of predictors at one time, the step-wise strategy will significantly reduce the computation burden. Moreover, by choosing the best bin set, we are able to control the model complexity yet keep prediction accuracy. Each scanning step will indicate us the best bin set, and the MARS model with this bin set is the model we are interested in. As we only consider the possible two-way interactions, we set the highest order of interaction to be two. Here we illustrate the model fitting procedure step by step.

- **First scanning step**

Suppose we have K bins. At each bin, fit a MARS model and calculate the Pearson's correlation coefficient r between observed and predicted expressions (Table 7). Therefore, we have K models. For the bth bin MARS model is fitted with following data matrix ($b = 1,2, \dots, K$).

	Response	Covariates for the First step			
		CF1	CF2	…	CFq
Gene 1	Y_1	$X1_b.1$	$X2_b.1$	…	$Xq_b.1$
Gene 2	Y_2	$X1_b.2$	$X2_b.2$	…	$Xq_b.2$
…	…	…	…	…	…
Gene n	Y_n	$X1_b.n$	$X2_b.n$	…	$Xq_b.n$

Table 7. Data for the first scan. At the first scanning step, the MARS model is fit to one features at one bin.

We mark the bin with largest Pearson's r value as the first best bin. The interaction term selected in this step would indicate the interaction between predictors within one position. The covariates in best model are marked as $X_{.1st}$ and brought into the second scan.

- **Second scanning step**

In this step, the model is fitted with features from two bins. One is the $X_{.1st}$ from the first scan, the other is one of the rest bins (Table 8). We fit $(K-1)$ models in this step. The new added bin with best prediction performance is marked as the second best bin. The data involved in this step would be:

	Response	First best bin				Covariates for the Second step			
		CF1	CF2	…	CFq	CF1	CF2	…	CFq
Gene 1	Y_1	$X1_{1st}.1$	$X2_{1st}.1$	…	$Xq_{1st}.1$	$X1_b.1$	$X2_b.1$	…	$Xq_b.1$
Gene 2	Y_2	$X1_{1st}.2$	$X2_{1st}.2$	…	$Xq_{1st}.2$	$X1_b.2$	$X2_b.2$	…	$Xq_b.2$
…	…	…	…	…	…	…	…	…	…
Gene n	Y_n	$X1_{1st}.n$	$X2_{1st}.n$	…	$Xq_{1st}.n$	$X1_b.n$	$X2_b.n$	…	$Xq_b.n$

Table 8. Data for the Second scan. At the second scanning step, the MARS model is fit to two bins. One of the two bins is the selected best bin from the first step and the other one is one of the rest bins.

The two-way interaction in this step may indicate the interaction between different covariates within same position, the interaction between different positions within the same covariate, and the interaction between different covariates at different positions. For example, the interaction terms of basis functions may have three forms:

$(X1_b - t_{X1_b})_+(X4_b - t_{X4_b})_+$ indicates the interaction between first feature and the forth feature at same position, the bth bin;

$(X1_{1st} - t_{X1_{1st}})_+(X1_b - t_{X1_b})_+$ indicates the interaction between first best bin and the bth bin within the first covariate ($b \neq 1st$);

and $(X1_{1st} - t_{X1_{1st}})_+(X4_b - t_{X4_b})_+$ indicates the interaction between first covariate in first best bin and the forth covariate in the bth bin.

The best model is chosen by Pearson's *r* value and the newly added bin in the best model is marked as the Second Best bin ($X_{.2nd}$). In the third step features from one of the rest K-2 bins will be modeled together with $X_{.1st}$ and $X_{.2nd}$. We repeat this process until get the satisfied model.

By doing so, our model is able to capture the comprehensive and spatial interactive effects among chromatin features. Although biologically it is still not clear that how this works through the epigenetic system, the contribution from interaction terms does suggest such potential interactive effects.

4 Real data analysis

4.1 Prediction performance and model complexity

For the real data, 1,373 genes (~10% of the total samples) were randomly sampled from the whole dataset as a test set, and the remaining genes were used in training models. This section shows the details of the model fitting at every scanning step and also the feature selection of the final model.

As the scanning went on, the prediction accuracy increased. We noticed that in each of the first five steps, only one bin would give the largest increment on the prediction accuracy (Table 9). At the sixth scanning, the Pearson's *r* reached a plateau and more than one bins achieved the same prediction performance. There was no unique best set of features could help improve the model fitting. At this point, the final model would stop at the fifth scanning step.

There were 14 chromatin features from bin 21^{st} , 28^{th} , 41^{st} ,24^{th} and 13^{th} , in total 70 features, involved in the feature selection for final model.

Steps of scanning	1	2	3	4	5	6
Pearson's *r* of the best model	0.9147	0.9287	0.9311	0.9339	0.9354	0.9354
Chosen bin ID at each step	21^{st}	28^{th}	41^{st}	24^{th}	13^{th}	33^{rd} 34^{th} 35^{th} 36^{th} 37^{th} 38^{th} 39^{th} 40^{th}

Table 9. Prediction performance and bins selected in each scanning step.

To avoid bias from the subsampling, we fitted the model with ten-fold cross validation. The whole data set was randomly split into ten equal size non-overlapping subsets. The model was trained by nine folds of the total sample and the rest one fold was used to validate the prediction performance (Table 10). After each step, the best models' Pearson's correlation coefficients (Pearson's *r*) were averaged across ten-fold cross-validation. Notice that, the model selection was specific to each of the ten folds.

Although the model complexity increased with adding new bins, the MARS backward model selection automatically kept good balance between prediction accuracy and model simplicity. With around 30 terms (included individual effect and two-way interactions), the model can obtain high correlation between fitted expressions and true expressions (Table 10).

With linear regression analysis, The Bestbin method in Greven et al. had Pearson's *r*=0.871 by only considering about the individual chromatin feature effect(Dong et al., 2012). In

our first scan, some interaction terms were selected. The average Pearson's *r* was 0.8912, which was better than the value computed from the Bestbin method. The selected model in first step actually showed significance on the interaction terms.

Steps of scanning	1	2	3	4	5	6
Average Pearson's *r* of the best model	0.8912	0.906	0.9125	0.9164	0.9184	0.9196
Average number of terms	24.1	24.4	24.3	27.5	28.8	28.9

Table 10. Average prediction performance of and model size by ten-fold cross-validation.

4.2 Chromatin features selection

The importance of each variable was calculated by evimp function in {earth} R package. It counted the number of model subsets that include a specific variable. Among the important features in the final model, five of them were selected from the first scanning step which was 21st bin. The most important feature was H3k79me2 in 28th bin (Figure 19). Lysine 79 of histone H3 (H3k79) can be mono-, di- or trimethylated by Dot1 methylase. The methylation at this residue acts as a marker of inactive chromatin regions that is critical for transcriptional silencing(Onder et al., 2012). The final model also indicated that the methylation of H3k4 was very important in gene regulation. According to the calculation, by adding H3k4me1 and H3k4me2, the generalized cross validation score showed a dramatic decrease (Figure 19). H3k4 methylation always associates with active transcription. Specifically, H3k4 di-methylation appears to be

related with the activation and potential activation of genes(Bernstein et al., 2002; Krogan et al., 2002; Ng et al., 2003).

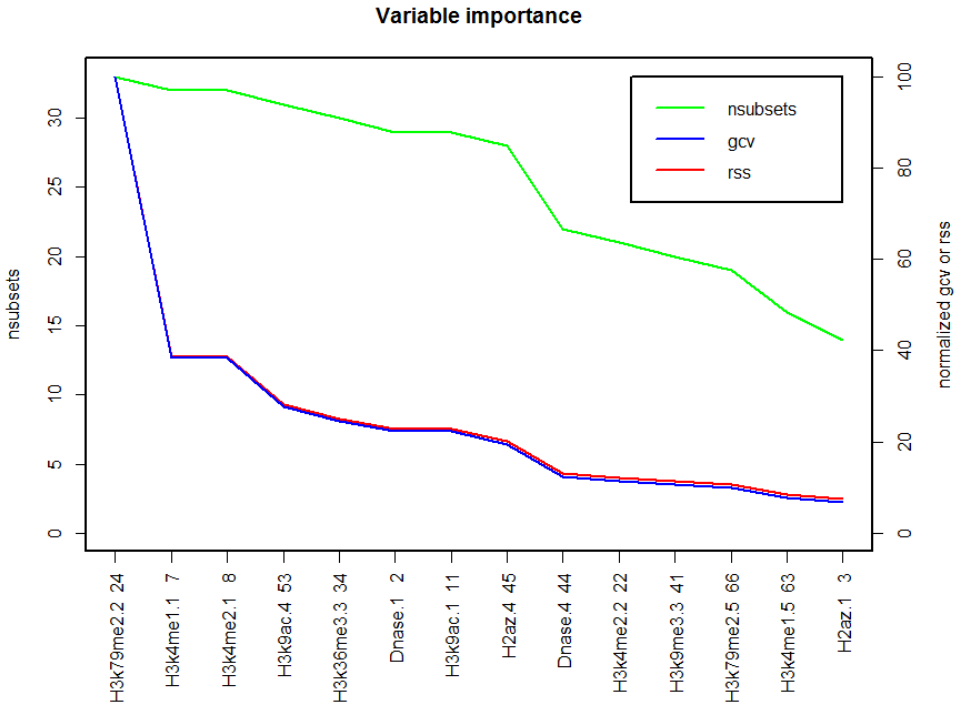


Figure 19. The plot for relative importance of chromatin features in final model. The importance of variables were judged by the number of sub-models (subsets) which include the specific predictor and the improvement of the model fit (decrease of GCV or RSS) by adding the predictor. The importance was ranked and plot in a descending order.

We collected the most important features shared by at least five of the ten models from ten-fold cross validation (Table 11). The subset of selected terms might vary by using different training data, but some common features did exist, for example, the promoter markers H3k4me2, H2A.Z, and H3k9ac. The di-methylation of histone H3 on lysine 4 (H3k4me2) was always associated with transcriptional activation. H3K9ac was also found in actively transcribed promoters (Koch et al., 2007). Histone H2A variant H2A.Z was associated with the promoters of

actively transcribed genes and also involves in the prevention of the spread of silent heterochromatin (Guillemette et al., 2005). Furthermore, H2A.Z had roles in chromatin for genome stability (Billon et al., 2012). These findings agree with the top important features from Bestbin model results. Our model indicated that, with information from more than one genomic locus, not all the chromatin features had to be used as the predictors.

	Important features
First scan	H3k4me1;H3k4me2;H3k79me2;H3k9ac;DnaseI;H3k27ac;H3k9me3;H3k4me3;H4k20me1;H2az;H3k9me1
Second scan	H3k4me2;H3k79me2;DnaseI;H2az;H3k4me1;H3k9ac;H3k9me3;H3k4me3;H4k20me1;H3k36me3
Third scan	H3k79me2;H3k4me2;DnaseI;H2az;H3k4me1;H3k9ac;H3k9me3;H3k36me3
Forth scan	H3k4me2;H3k79me2;DnaseI;H3k4me1;H2az;H3k9ac;H3k36me3;H3k9me3
Fifth scan	H3k79me2;H3k4me2;DnaseI;H2az;H3k4me1;H3k9ac;H3k9me3;H3k36me3

Table 11. Overlapped important features by ten-fold cross-validation. The listed chromatins were most common selected in the ten-fold cross-validation at each scanning step.

In order to check if different subgroups of chromatin features may act similarly as previous studies (Dong et al., 2012), we also fitted our model to different categories of chromatin features. The promoter marks, structural marks, repressive marks and distal/other marks together with the combination of groups.

The average Pearson's *r* values, obtained by average across ten-fold cross-validation, show that repressive and distal markers are least informative in predicting gene expressions

(Figure 20, Table S 2). The most predictive marks were promoter marks and all combinations of subset marks that include promoter marks perform equally well in prediction. This was partially because of the CAGE TSS-based expression profiles were more sensitive to capture the transcription initiation events. Although this was in agreement with the result from Bestbin method (Dong et al., 2012), by repressive makers alone, the prediction performance of our method (Pearson's $r = 0.632$) was much better (maximum Pearson's $r = \sim 0.5$). Overall, the prediction accuracy was improved by our stepwise molding and incorporation of interaction terms.

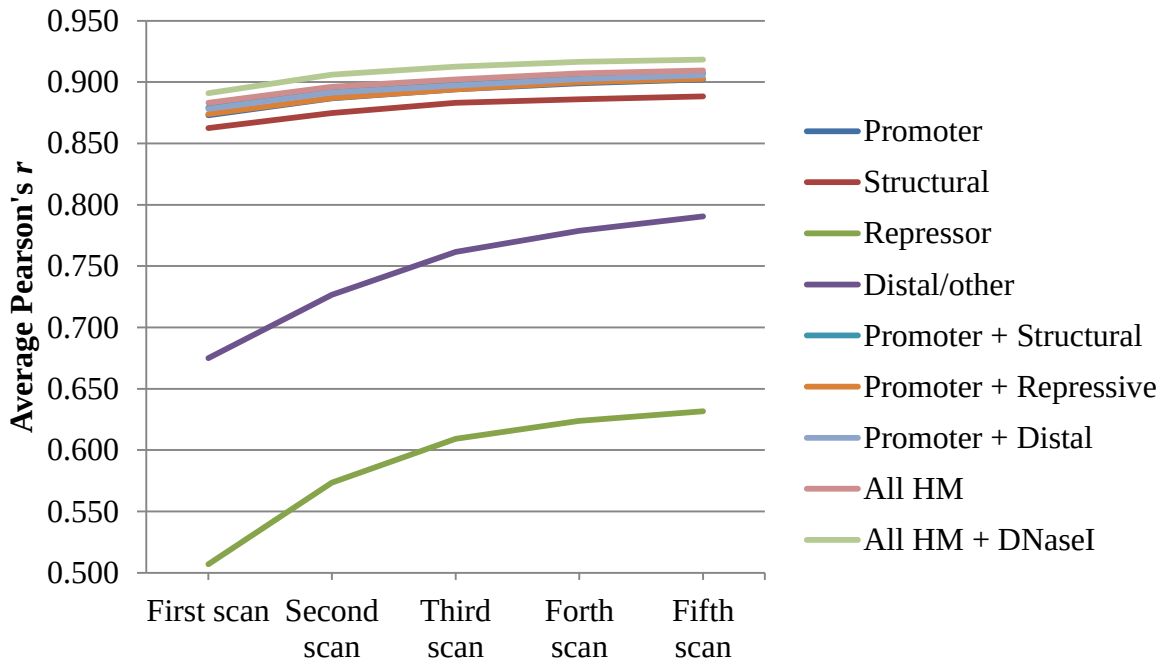


Figure 20. Average Pearson's *r* values for prediction with different subset of chromatin features. Promoter marks: H3K4me2, H3K4me3,H2A.Z, H3K9ac and H3K27ac ; structural marks: H3K36me3 and H3K79me2; repressive marks: H3K27me3 and H3K9me3and distal/other marks: H3K4me1, H4K20me1 and H3K9m (Raisner et al., 2005; Barski et al., 2007; Benevolenskaya 2007; Koch et al., 2007; Steger et al., 2008; Kolasinska-Zwierz et al., 2009).

4.3 Spatial component of chromatin features

From the figure showing the Pearson's r on every scan (Figure 21), we noticed that the prediction accuracy of model has an obvious peak at the center (the TSS site) in first scan. In selection of the bin by model fitting, the closer a bin was to the TSS (20^{th} bin) the more it would contribute to the gene expression. In our one-run model, the first scan chose the 21^{st} bin (the bin located at the TSS) as the first best bin (Figure 21). This is consistent with the previous study, which indicated the bin close to the TSS would be more informative in prediction (Cheng, Yan et al. 2011). Secondly, although the Pearson's r values were relatively higher at the center (TSS), after the first scan the Pearson's r values increased to around 0.92 and when adding new bin the values across all bins were almost similar.

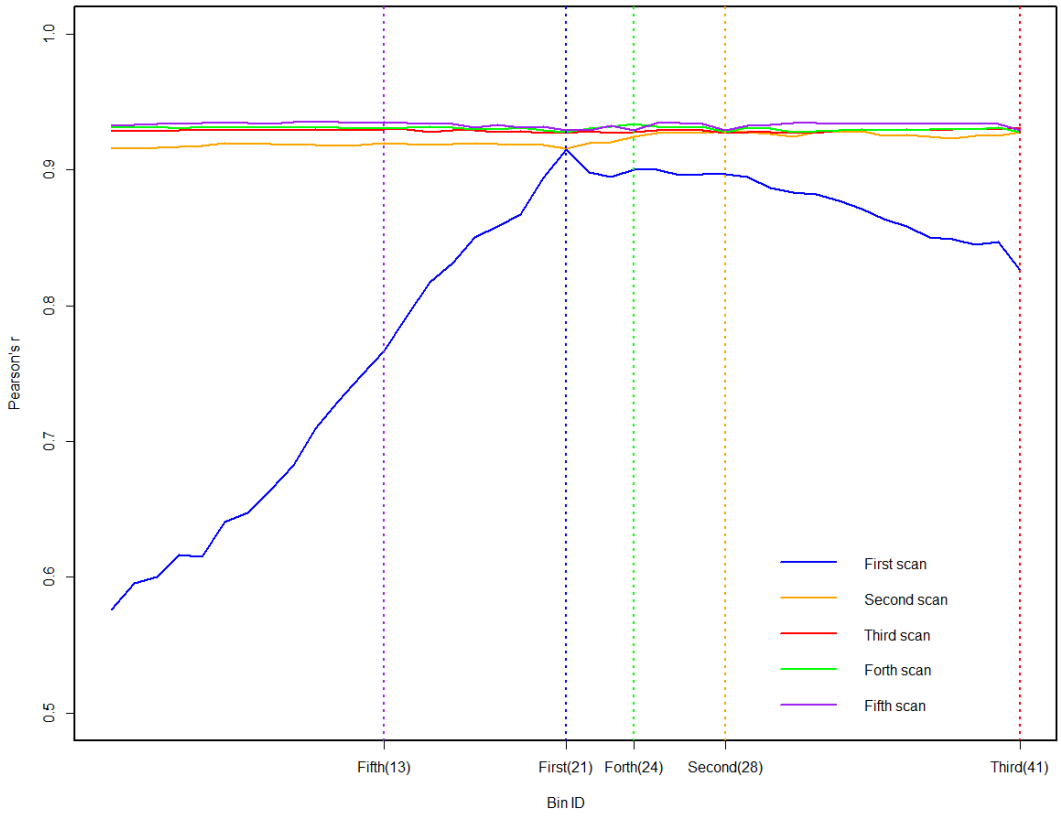


Figure 21. Pearson's r values for each scanning step. The vertical dotted lines denote the largest values for each scanning step which indicate the selected bins. The obvious peak at the TSS (21st bin) disappeared by adding the second best bin in the model. The Pearson's *r* increased with the step-wise scanning but the pace improvement decreased.

Interestingly, the genebody bin (41st bin) was chosen in the third scan and retained in the final model (Table S 3). We found some interaction terms for modifications at the TSS with that in the gene body (Table S 3). Histone modification marks positioned in the gene bodies may have role in determining the splicing patterns. Recent studies have noted that the abundance of nucleosomes might count for the observed exonic gene body histone modification marks. It was claimed that the positioned nucleosomes at exons might enhance splicing by increasing RNAPII occupancy time (Kornblihtt et al., 2009). In our model, it suggested the effect from gene body

marks contributed combinationally with the marks at the TSS region. Such interactions between different genome loci may globally exist given the complexity of the transcriptional regulation mechanism. The detail of the interaction terms will be discussed in section 4.4.

Our ten-fold cross-validation showed the bin chosen results are not sampling dependent (Table 12). The 21st was always picked as the crucial position and there was a strong indication of the systematically contribution from genebody (41st bin) (Table 12).

Fold ID	1	2	3	4	5	6	7	8	9	10
First scan	21st	21st	21st	21st	23rd	21st	26th	21st	21st	21ts
Second scan	41st	26th	28th	30th	20th	30th	20th	31st	20th	27th
Third scan	22nd	20th	16th	20th	21st	22nd	22nd	22nd	27th	22nd
Forth scan	19th	19th	41st	24th	29th	34th	41st	41st	41ts	41st
Fifth scan	28th	34th	22nd	41st	17th	17th	18th	17th	19th	11th

Table 12. The bin selected in each step by each fold of the ten-fold cross-validation. Nine out of ten folds chose the 21st bin and seven folds chose the genebody bin (41st bin). This indicated the genebody effect was not bias to subsampling. It also suggested the interactions between chromatin features in genebody and that in the TSS region.

4.4 Interactions between chromatin features

Since the evidence firstly presented that the functional significance of histone modifications, it has been established the reasonable doubt that specific modifications and combinations would mediate protein-protein interactions crucial for the translational regulation. Some of the interactions involved how particular histone tail modifications can interact. In the past few years, a lot of studies attempted to find the support to the existence of the histone code. This is especially meaningful when a histone code could be heritable and involve in a long-term maintenance of a transcriptional state.

4.4.1 Interactions between chromatin variants and histone modifications

There were 14 features chosen from 70 (14 features from 5 bins) in the final fitted model (Table S 3). The model included 37 terms after forward stepwise fitting and the final model maintained 34 of them. Among the 33 basis functions (excluding the intercept term), 13 were main effects and 20 were interaction terms. Among all the selected interaction terms, it indicated that the histone variants Dnase I and H2A.Z were found interacted with other histone modifications (Table 13). It suggested that some histone modification may impact the gene expressions differently under various densities of histone variants signals.

The DNase I hypersensitive sites are regions that are sensitive to cleavage by enzyme Dnase I. Such regions tend to lose the condense structure and therefore exposing the DNA. These accessible genome regions are functionally related to transcriptional activity. (Thurman et

al., 2012). A recent quantitative study demonstrated that the genome-wide combinational effects of chromatin features to differential Dnase I hypersensitivity. It was observed that the methylation H3k4 and acetylation of both H3k9 and H3k27 were sharply elevated in different types of DNaseI hypersensitivity (DNaseI HS) sites in K562 cell type (Shu et al., 2011). In the set of interactions of our final model, Dnase I signal in bin#21 interacted with the H3k9ac signal in both bin#21 and bin#24 (Table 13). It also showed the potential interactions between Dnase I and H3k4me2. In addition to the simply positive or negative correlation between DNaseI HS and histone modification, our study uncovered the different quantitative effect of histone modification on gene expression under varieties of DNaseI enrichment. For example, the critical effect that H3k9ac has for transcriptional elongation may vary according to the different levels of DNaseI HS.

It was revealed that the combinational presence of H2A.Z at active gene promoters with at least three specific H3 acetylation sites at k9, k18 and k27 associated with the gene activation (Raisner et al., 2005; Wang et al., 2008). ChIP-seq studies in humans have indicated that H2A.Z is localized to gene enhancers and promoters with a positive correlation between occupancy and transcriptional activity at promoters. It was also suggested that H2A.Z might function to increase nucleosome mobility by destabilizing nucleosome structure. It was reported that H2A.Z ad H3k27me3 are colocalized at low-expressed genes in embryonic stem cells (Creyghton et al., 2008).Another study also indicated that the colocalization of H2A.Z with H3k9ac and H3k4me3 suggested that these histone marks were perfectly deposited on H2A.Z enriched nucleosomes(Bártfai et al., 2010). Such interactions may support the model in which the H2A.Z enriched nucleosomes serve to demarcate regulatory regions in the genome and promote transcription initiation by guiding chromatin modifying and transcription initiating.

Basis functions	Parameter estimations
h(Dnase.1-4.93013) * h(H3k4me2.1-5.04162)	0.703
h(Dnase.1-4.93013) * h(5.04162-H3k4me2.1)	-0.788
h(4.93013-Dnase.1) * h(H3k9ac.1-4.16879)	-0.415
h(Dnase.1-0.23396) * h(3.86846-H3k79me2.2)	-0.069
h(0.23396-Dnase.1) * h(3.86846-H3k79me2.2)	0.153
h(4.93013-Dnase.1) * h(Dnase.4-2.0571)	-0.230
h(Dnase.1-2.83178) * h(H3k9ac.4-1.51068)	-0.141
h(2.83178-Dnase.1) * h(H3k9ac.4-1.51068)	-0.053
h(H2az.1-4.85662) * h(H3k9ac.4-1.51068)	-0.197
h(4.85662-H2az.1) * h(H3k9ac.4-1.51068)	-0.095

Table 13. Interaction terms between histone variants Dnase I /H2A.Z with histone modifications in final model. The number ".1", ".2" and ".4" indicate the 21st bin, 28th bin and 24th bin selected by scanning step respectively.

4.4.2 Interactions between histone modifications

A few recent studies discussed about the "bivalent domains" of the histone modifications (Bernstein et al., 2006; Wang et al., 2008). It has been proposed that such colocalization of multiple epigenetic modifications plays an important role in the gene regulation.

The genome wide study of human CD4+ T cells detected the "backbone" modification module consisting of 17 (H2A.Z, H2BK5ac, H2BK12ac, H2BK20ac, H2BK120ac,H3K4ac, H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me1, H3K18ac, H3K27ac, H3K36ac, H4K5ac, H4K8ac and H4K91ac) modifications at 3286 promoters (Wang et al., 2008). The study

suggested that these modifications tend to colocalize at gene promoters and most of them were found correlated with each other at an individual nucleosome level. This suggested the functional complexity of interplay between modifications. Another high-resolution profiling histone methylation study revealed the bivalent domains with both H3k4me3 and H3k27me3 signals exist next to each other. The H3k4me3 signals were elevated at the promoter while H3k27me3 signals were distributed over broader regions. Such dynamic pattern was suggested to play regulatory roles for the differentiation of embryonic stem cells (Bernstein et al., 2006; Barski et al., 2007). The existing studies proposed the occupancy of combination of histone modifications were functional in gene regulation, while they always focused on the peaks of the signals. The quantitative association of such epigenetic complexity with the gene regulation was largely unknown.

Our study clearly indicated the existence of such interactions. In the same bin at TSS (21st bin) we observed interactions between H3k4me1 and H3k4me2 as well as between H3k4me1 and H3k9ac (Table 14). The H3k4 methylation were always found elevated at the TSS regions. Both H3k4me1 and H3k4me2 are positively related to the transcriptional levels. In the interaction between H3k4me1 and H3k4me2 at the same bin (21st bin), we find that under a threshold value of H3k4me1 the marginal effect from H3k4me2 is slightly varying (Table 14). Similar combinational effect was found for the H3k9ac, which is the essential activation mark near the TSS.

As discussed in section 4.3, we identified a set of interactions between modifications located in genebody and that located in TSS regions. For example, the interaction between gene body mark H3k36me3 and the activation mark H3k9ac at TSS (Table 14). The H3k36me3 always is found highly occupied after TSS in active regions. However, the interaction suggested

that this effect may differ if H3k9ac signals at the TSS reached at some specific threshold value. The interactions between H3k9me3 and H3k9ac were also worth noticing. The signals of H3k9me3 were known to be higher in TTS regions of silent genes. Recent study also proposed that the genes with H3k9me3 in gene body are associated with low levels of transcription (Hahn et al., 2011). The interactive relationship between repression mark at gene body and the activation mark at the TSS is an evidence to support the hypothesis of histone code.

Basis functions	Parameter estimations
h(3.67525-H3k4me1.1) * h(H3k4me2.1-1.80456)	0.456
h(3.67525-H3k4me1.1) * h(1.80456-H3k4me2.1)	-0.368
h(3.67525-H3k4me1.1) * h(H3k9ac.4-2.85807)	-0.200
h(3.67525-H3k4me1.1) * h(2.85807-H3k9ac.4)	-1.092
h(H3k36me3.3-1.17832) * h(H3k9ac.4-1.51068)	0.165
h(1.17832-H3k36me3.3) * h(H3k9ac.4-1.51068)	-0.767
h(H3k9me3.3-0.800334) * h(H3k9ac.4-1.51068)	-0.319
h(0.800334-H3k9me3.3) * h(H3k9ac.4-1.51068)	0.106

Table 14. Interaction terms between histone modifications in final model. The activation mark H3k9ac at TSS was find interacted with the activation mark H3k36me3 and repression mark H3k9me3 at genebody bin. The number ".1", ".3" and ".4" indicate the 21st bin, 41st bin and 24th bin selected by scanning step respectively.

5 Simulation study

To compare the prediction performance of our SMARS with Bestbin method, we did three sets of simulation studies with different generation functions. We chose five real chromatin features as covariates and three bins from each. So for covariates we had $(X1_1, X1_2, X1_3, \dots\dots, X5_1, X5_2, X5_3)$. The error term was a random variable generated from normal distribution with mean 0 and variance 1. For all model evaluation, 10,000 samples were randomly chosen as the training set, 3,731 samples were for testing the prediction performance.

5.1 Linear function with one bin per covariate

In the first simulation setting, response values were generated by a simple linear function and only one bin from each variable was included.

$$y = 5 + 10x1_1 + x2_2 + 4x3_3 + 2x4_2 + error$$

Bestbin method chose the right bin through correlation except X2. The parameters estimations were quite satisfied with high significance. The Bestbin final model was

$$f(x) = 4.18 + 9.97x1_1 + 0.88x2_1 + 3.98x3_3 + 2.11x4_2$$

The SMARS method chose the bin1, bin3 and bin2 at three scan steps respectively. We found the difference of models at the second step and the third step showed no significant improvement. This is due to the major variance of the response could be explained with the

information from bin1 and bin3. This is consistent with the best bin model which indicates the least significance of term $x4_2$. The final SMARS model was

$$f(x) = 26.37 + 9.92(x1_1 - 1.54)_+ - 10(1.54 - x1_1)_+ + 3.49(x4_2 - 3.2)_+ - 2.14(3.2 - x4_2)_+ + 3.95(\mathrm{x}3_3 - 0.7)_+ - 4.3(0.7 - x3_3)_+$$

Two models had quite similar prediction accuracy, as the Pearson's *r* was 0.9984 vs. 0.9983 (Figure 22). This indicates that when underlying true model is a regular simple linear model, both feature pre-selection of best bin and the stepwise feature selection of SMARS can work equally well in both prediction and identifying true model.

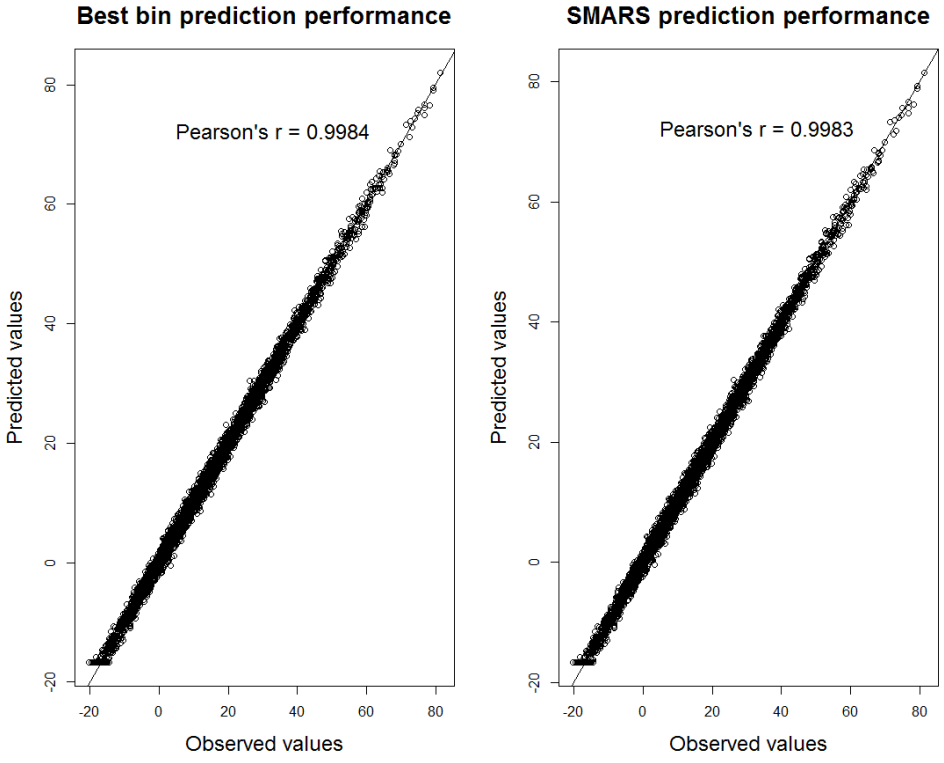


Figure 22. Scatter plot of predicted response values versus observed values with linear generation function. Both Bestbin method and SMARS achieved similar good prediction accuracy.

5.2 Linear function with multiple bins per covariate in additive way

The second linear generation function incorporated at least two bins from each feature.

$$y = 5 + 3x1_1 + 4x1_3 + 10x2_2 + 10x2_3 + 6x3_1 + 20x3_2 + 15x3_3 + 8x4_1 + 7x4_2 + x5_1 + 15x5_2 + \text{error}$$

The fitted function from Bestbin model identified the represented bin for each feature. The prediction was good as the bin selection chose the variable which had larger impact on the response values.

$$f(x) = -8.84 + 13.56x1_3 + 16.07x2_2 + 29.02x3_2 + 12x4_2 + 18.8x5_2$$

The SMARS chose bin2, bin3 and bin1 to add into the model at each scan. Ten out of fifteen predictors were selected. Although involved more parameters than Bestbin method, the SMARS model did not lose the adaption to the test data set and get a slightly better prediction (Figure 23).

$$\begin{aligned} f(x) = 148.5 &+ 3.2(x1_1 - 3.58)_+ - 3.04(3.58 - x1_1)_+ + 4(x1_3 - 1.17)_+ \\ &- 4.06(1.17 - x1_3)_+ + 10.02(x2_2 - 1.62)_+ - 10.24(1.62 - x2_2)_+ \\ &+ 10.62(4.01 - x2_3)_+ - 10.03(x2_3 - 4.01)_+ + 6.04(x3_1 + 1.49)_+ \\ &- 6.04(-1.49 - x3_1)_+ + 19.98(x3_2 - 1.18)_+ - 20.01(1.18 - x3_2)_+ \\ &+ 15(x3_3 + 0.7)_+ - 15(-0.7 - x3_3)_+ + 8.42(x4_1 - 2.27)_+ \\ &- 8.05(2.27 - x4_1)_+ + 7.03(x4_2 - 0.94)_+ - 7.02(0.94 - x4_2)_+ \\ &+ 15.25(x5_2 - .71)_+ - 15.07(2.71 - x5_2)_+ \end{aligned}$$

Under this generation function, as the Bestbin can catch the major part of variance, it was able to yield satisfied prediction accuracy. While in the favor of identifying the true model, the SMARS gave us more information for the underlying generation function.

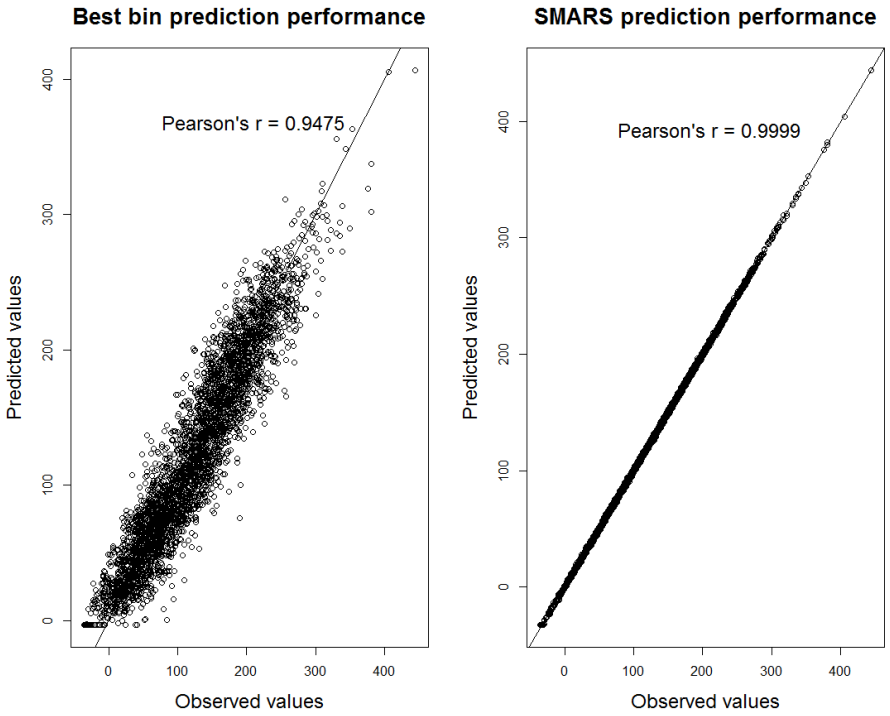


Figure 23. Scatter plot of predicted response values versus observed values with linear generation function included multiple measurements per feature. Although Bestbin method can get similar prediction performance to SMARS, it is not as powerful as SMARS in detecting the true function.

5.3 Interacting functions with multiple bins per covariate in productive way (including nonlinear terms)

In this setting, we included the interaction term by multiplying the features and we also created the nonlinear terms.

$$y = 5 + 3X1_1 + 20\sin(X1_1X3_1 \cdot \pi) + 3X2_2 + 8X3_1 + 15X2_2X3_3 + 2X3_3 - 15X2_3X4_1 + 7X4_1 + 8X4_2{}^2 + error$$

The representative best bin feature selected by Bestbin method was no longer adequate to approximate the true function.

$$f(x) = 3.3 + 2.12x1_3 + 11.52x2_2 + 33.29x3_3 - 3.32x4_3 - 2.87x5_3$$

In the final model with SMARS, the interactions $x2_2x3_3$ and $x2_3x4_1$ were selected. Three terms for $x4_2$ formed the quadratic effect. Moreover, the nonlinear interaction term $20\sin(x1_1x3_1 \cdot \pi)$ was fitted by four productions of basis functions of $x1_1$ and $x3_1$. Obviously, under generation setting, the SMARS was more appropriate for both prediction and interpretation (Figure 24).

$$\begin{aligned} f(x) = {} & -85.84 + 8.52(x3_1 + 1.49)_+ - 18.15(-1.49 - x3_1)_+ - 28(x4_1 - 2.07)_+ \\ & + 21.33(2.07 - x4_1)_+ - 19.02(x2_2 - 3.23)_+ + 10.16(3.23 - x2_2)_+ \\ & + 17.57(x4_2 + 0.087)_+ + 9.1(x4_2 - 0.96)_+ + 10.89(0.96 - x4_2)_+ \\ & - 31.88(x2_3 - 1.7)_+ + 61.76(x3_3 + 0.7)_+ \\ & - 28.52(-0.7 - x3_3)_+ 5.26(x1_1 + 1.29)_+(-1.49 - x3_1)_+ \\ & - 86.88(-1.29 - x1_1)_+(-1.49 - x3_1)_+ + 0.38(x1_1 - 0.31)_+(x3_1 + 1.49)_+ \\ & - 1.99(0.31 - x1_1)_+(x3_1 + 1.49)_+ + 15.35(2.07 - x4_1)_+(x2_3 - 1.9)_+ \\ & - 15.2(2.07 - x4_1)_+(1.9 - x2_3)_+ + 25.22(x2_2 - 3.81)_+(x3_3 + 0.7)_+ \\ & - 16.17(3.81 - x2_2)_+(x3_3 + 0.7)_+ \end{aligned}$$

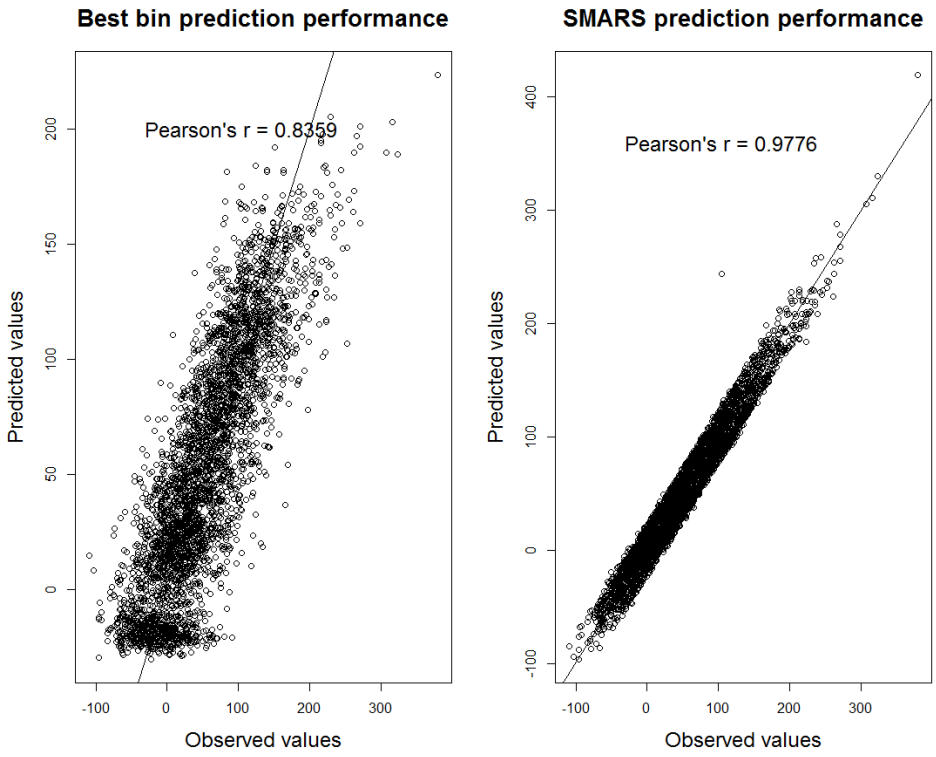


Figure 24. Scatter plot of predicted response values versus observed values with interacting generation function. The prediction accuracy of Bestbin method was not as good as SMARS.

Our simulation studies indicated that, if a simpler linear function truly existed, both methods can be similarly satisfied. While with a potential complexity of the underlying true function include interactions and non-linearity, SMARS is a better approach to approximate the association.

6 Conclusions

In this study, we proposed a modified MARS (S-MARS) model which could be adapted to the data that has multiple measurements for each feature. By step-wise fitting procedure, our method could incorporate the two-way interactions among the features as well as across different measurement points.

We applied SMARS model for predicting TSS-based gene expressions with the chromatin feature enrichment. Compared to the up-to-date existed method (Pearson's $r = 0.87$), we successfully improved the prediction accuracy (average Pearson's $r = 0.92$). Our model also improved the model interpretation which helped to further explore the complexity of epigenetic gene regulation mechanism as well as to support the hypothesis of histone code. Our finding indicated that, the combinational effect of histone modifications may exist in the epigenetic gene regulation. In addition to that, the epigenetic modifications at different genome coordinates may contribute corporately in gene activation or repression. Interestingly, we identified the interaction between the modification signals at the gene body and that at the TSS. This has not been uncovered by the past quantitative correlation studies.

With the three sets of simulation studies, we generalized the application of SMARS to different scenarios of generation functions from simple linear function to the complicated interactive and nonlinear function. It was shown that, comparing to the Bestbin method which would lose the information from multiple measurements, our SMARS model could have broader adaption to approximate both the simple and complicated functions. More importantly, in the sense of identifying the true generation function, SMARS is more likely to yield an informative model with advantage in interpretability.

Chapter IV: Discussion and future work

Both DNA methylation and histone modification have their own roles in establishing patterns of gene repression during development. Although DNA methylation and histone modification are carried out by different chemical reactions and require different enzymes, it is believed that there exists a *bona fide* connection between them. For example, the presence of DNA methyl groups can effect histone modification in a process that might be mediated by methyl binding proteins(Cedar et al., 2009). Recently, genetic links between histone methylation, notably H3k9me3, and DNA methylation have appeared in organisms as diverse as fungi, plants, and mice (Hashimshony et al., 2003; Lehnertz et al., 2003). Thus, histone modifications and DNA methylation are directly and indirectly connected to each other, contributing to the readout of higher-order chromatin structures.

There are still many mechanistic details of these inter-connections that need to be clarified. For example, it is known that the presence of methyl groups in DNA may affect chromatin packaging, but it is still unknown how the DNA methylation pattern is translated to produce the corresponding histone modification profile. Furthermore, it is barely understood about how the formation of histone methylation patterns may affect the *de novo* DNA methylation. Understanding the relationship between DNA methylation and certain histone modifications will provide insights into the abnormal gene expression patterns observed in diseases especially cancer. Studies have shown that the cancer cells are subject to aberrant *de novo* DNA methylation, and there is evidence suggests that this process may be related to certain histone modifications.

Besides the DNA methylation and histone modifications, there exist other factor that influence the epigenome, such as the three-dimensional chromatin architecture, noncoding RNA and protein binding in relation to chromatin modifications. Current studies support that these factors are not independent elements of functional epigenomes. With the DNA microarray based techniques and high-throughput sequencing techniques, there is greater potential to unveil the epigenetic regulation.

References

Banerjee, J., R. Magnani, M. Nair, L. M. Dirk, S. DeBolt, I. B. Maiti and R. L. Houtz (2013). "Calmodulin-mediated signal transduction pathways in Arabidopsis are fine-tuned by methylation." The Plant Cell Online **25**(11): 4493-4511.

Barski, A., S. Cuddapah, K. Cui, T. Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev and K. Zhao (2007). "High-resolution profiling of histone methylations in the human genome." Cell **129**(4): 823-837.

Bártfai, R., W. A. Hoeijmakers, A. M. Salcedo-Amaya, A. H. Smits, E. Janssen-Megens, A. Kaan, M. Treeck, T.-W. Gilberger, K.-J. Françoijs and H. G. Stunnenberg (2010). "H2A. Z demarcates intergenic regions of the Plasmodium falciparum epigenome that are dynamically marked by H3K9ac and H3K4me3." PLoS pathogens **6**(12): e1001223.

Becker, C., J. Hagmann, J. Müller, D. Koenig, O. Stegle, K. Borgwardt and D. Weigel (2011). "Spontaneous epigenetic variation in the Arabidopsis thaliana methylome." Nature **480**(7376): 245-249.

Benevolenskaya, E. V. (2007). "Histone H3K4 demethylases are essential in development and differentiation This paper is one of a selection of papers published in this Special Issue, entitled 28th International West Coast Chromatin and Chromosome Conference, and has undergone the Journal's usual peer review process." Biochemistry and cell biology **85**(4): 435-443.

Bernstein, B. E., E. L. Humphrey, R. L. Erlich, R. Schneider, P. Bouman, J. S. Liu, T. Kouzarides and S. L. Schreiber (2002). "Methylation of histone H3 Lys 4 in coding regions of active genes." Proceedings of the National Academy of Sciences **99**(13): 8695-8700.

Bernstein, B. E., T. S. Mikkelsen, X. Xie, M. Kamal, D. J. Huebert, J. Cuff, B. Fry, A. Meissner, M. Wernig and K. Plath (2006). "A bivalent chromatin structure marks key developmental genes in embryonic stem cells." Cell **125**(2): 315-326.

Bestor, T. H. (2000). "The DNA methyltransferases of mammals." Human molecular genetics **9**(16): 2395-2402.

Billon, P. and J. Côté (2012). "Precise deposition of histone H2A. Z in chromatin for genome expression and maintenance." Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms **1819**(3): 290-302.

Braunstein, M., A. B. Rose, S. G. Holmes, C. D. Allis and J. R. Broach (1993). "Transcriptional silencing in yeast is associated with reduced nucleosome acetylation." Genes Dev **7**(4): 592-604.

Breiman, L., J. H. Friedman, R. A. Olshen and C. J. Stone (1984). "Classification and regression trees. Wadsworth & Brooks." Monterey, CA.

Burn, J., D. Bagnall, J. Metzger, E. Dennis and W. Peacock (1993). "DNA methylation, vernalization, and the initiation of flowering." Proceedings of the National Academy of Sciences **90**(1): 287-291.

Cedar, H. and Y. Bergman (2009). "Linking DNA methylation and histone modification: patterns and paradigms." Nature Reviews Genetics **10**(5): 295-304.

Chan, S. W.-L., I. R. Henderson and S. E. Jacobsen (2005). "Gardening the genome: DNA methylation in Arabidopsis thaliana." Nature Reviews Genetics **6**(5): 351-360.

Chan, S. W.-L., D. Zilberman, Z. Xie, L. K. Johansen, J. C. Carrington and S. E. Jacobsen (2004). "RNA silencing genes control de novo DNA methylation." Science **303**(5662): 1336-1336.

Chan, T. L., S. T. Yuen, C. K. Kong, Y. W. Chan, A. S. Chan, W. F. Ng, W. Y. Tsui, M. W. Lo, W. Y. Tam, V. S. Li and S. Y. Leung (2006). "Heritable germline epimutation of MSH2 in a family with hereditary nonpolyposis colorectal cancer." Nat Genet **38**(10): 1178-1183.

Cheng, C., K. K. Yan, K. Y. Yip, J. Rozowsky, R. Alexander, C. Shou and M. Gerstein (2011). "A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets." Genome Biol **12**(2): R15.

Cheng, X. and R. M. Blumenthal (2008). "Mammalian DNA methyltransferases: a structural perspective." Structure **16**(3): 341-350.

Chiang, I. C., G. C. Liu, R. S. Sheu and Y. T. Kuo (2001). "High resolution CT of Wegener's granulomatosis--case reports." Kaohsiung J Med Sci **17**(4): 221-225.
Chin, D. and A. R. Means (2000). "Calmodulin: a prototypical calcium sensor." Trends in cell biology **10**(8): 322-328.
Cokus, S. J., S. Feng, X. Zhang, Z. Chen, B. Merriman, C. D. Haudenschild, S. Pradhan, S. F. Nelson, M. Pellegrini and S. E. Jacobsen (2008). "Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning." Nature **452**(7184): 215-219.
Craven, P. and G. Wahba (1978). "Smoothing noisy data with spline functions." Numerische Mathematik **31**(4): 377-403.
Creyghton, M. P., S. Markoulaki, S. S. Levine, J. Hanna, M. A. Lodato, K. Sha, R. A. Young, R. Jaenisch and L. A. Boyer (2008). "H2AZ is enriched at polycomb complex target genes in ES cells and is necessary for lineage commitment." Cell **135**(4): 649-661.
Cuthbert, G. L., S. Daujat, A. W. Snowden, H. Erdjument-Bromage, T. Hagiwara, M. Yamada, R. Schneider, P. D. Gregory, P. Tempst and A. J. Bannister (2004). "Histone deimination antagonizes arginine methylation." Cell **118**(5): 545-553.
Dolinoy, D. C. and R. L. Jirtle (2008). "Environmental epigenomics in human health and disease." Environ Mol Mutagen **49**(1): 4-8.
Dong, X., M. C. Greven, A. Kundaje, S. Djebali, J. B. Brown, C. Cheng, T. R. Gingeras, M. Gerstein, R. Guigó and E. Birney (2012). "Modeling gene expression using chromatin features in various cellular contexts." Genome Biol **13**(9): R53.
Donoho, D. L. and I. M. Johnstone (1989). "Projection-based approximation and a duality with kernel methods." The Annals of Statistics: 58-106.
Egger, G., G. Liang, A. Aparicio and P. A. Jones (2004). "Epigenetics in human disease and prospects for epigenetic therapy." Nature **429**(6990): 457-463.
Feinberg, A. P. and B. Tycko (2004). "The history of cancer epigenetics." Nature Reviews Cancer **4**(2): 143-153.
Feinberg, A. P. and B. Tycko (2004). "The history of cancer epigenetics." Nat Rev Cancer **4**(2): 143-153.
Finnegan, E., R. Genger, K. Kovac, W. Peacock and E. Dennis (1998). "DNA methylation and the promotion of flowering by vernalization." Proceedings of the National Academy of Sciences **95**(10): 5824-5829.
Finnegan, E. J., W. J. Peacock and E. S. Dennis (1996). "Reduced DNA methylation in Arabidopsis thaliana results in abnormal plant development." Proceedings of the National Academy of Sciences **93**(16): 8449-8454.
Friedman, J. H. (1991). "Multivariate adaptive regression splines." The annals of statistics: 1-67.
Friedman, J. H., E. Grosse and W. Stuetzle (1983). "Multidimensional additive spline approximation." SIAM Journal on Scientific and Statistical Computing **4**(2): 291-301.
Friedman, J. H. and W. Stuetzle (1981). "Projection pursuit regression." Journal of the American statistical Association **76**(376): 817-823.
Goll, M. G. and T. H. Bestor (2005). "Eukaryotic cytosine methyltransferases." Annu. Rev. Biochem. **74**: 481-514.
Gordon, A. and G. Hannon (2010). "Fastx-toolkit." FASTQ/A short-reads pre-processing tools (unpublished) http://hannonlab. cshl. edu/fastx_toolkit.
Guillemette, B., A. R. Bataille, N. Gévry, M. Adam, M. Blanchette, F. Robert and L. Gaudreau (2005). "Variant histone H2A. Z is globally localized to the promoters of inactive yeast genes and regulates nucleosome positioning." PLoS biology **3**(12): e384.
Hahn, M. A., X. Wu, A. X. Li, T. Hahn and G. P. Pfeifer (2011). "Relationship between gene body DNA methylation and intragenic H3K9me3 and H3K36me3 chromatin marks." PLoS One **6**(4): e18844.

Hashimshony, T., J. Zhang, I. Keshet, M. Bustin and H. Cedar (2003). "The role of DNA methylation in setting up chromatin structure during development." Nature genetics **34**(2): 187-192.
Hassa, P. O., S. S. Haenni, M. Elser and M. O. Hottiger (2006). "Nuclear ADP-ribosylation reactions in mammalian cells: where are we today and where are we going?" Microbiology and Molecular Biology Reviews **70**(3): 789-829.
Hastie, T., R. Tibshirani, J. Friedman, T. Hastie, J. Friedman and R. Tibshirani (2009). The elements of statistical learning, Springer.
Hedhly, A., J. I. Hormaza and M. Herrero (2009). "Global warming and sexual plant reproduction." Trends in plant science **14**(1): 30-36.
Hoang, S. A., X. Xu and S. Bekiranov (2011). "Quantification of histone modification ChIP-seq enrichment for data mining and machine learning applications." BMC research notes **4**(1): 288.
Holliday, R. and J. E. Pugh (1975). "DNA modification mechanisms and gene activity during development." Science **187**(4173): 226-232.
Jenuwein, T. and C. D. Allis (2001). "Translating the histone code." Science **293**(5532): 1074-1080.
Jenuwein, T. and C. D. Allis (2001). "Translating the histone code." Science **293**(5532): 1074-1080.
Jones, P. A. and S. M. Taylor (1980). "Cellular differentiation, cytidine analogs and DNA methylation." Cell **20**(1): 85-93.
Kakutani, T., J. A. Jeddeloh and E. J. Richards (1995). "Characterization of an Arabidopsis thaliana DNA hypomethylation mutant." Nucleic acids research **23**(1): 130-137.
Karlic, R., H. R. Chung, J. Lasserre, K. Vlahovicek and M. Vingron (2010). "Histone modification levels are predictive for gene expression." Proc Natl Acad Sci U S A **107**(7): 2926-2931.
Kato, M., A. Miura, J. Bender, S. E. Jacobsen and T. Kakutani (2003). "Role of CG and Non-CG Methylation in Immobilization of Transposons in< i> Arabidopsis</i>." Current Biology **13**(5): 421-426.
Kent, W. J., A. S. Zweig, G. Barber, A. S. Hinrichs and D. Karolchik (2010). "BigWig and BigBed: enabling browsing of large distributed datasets." Bioinformatics **26**(17): 2204-2207.
Kim, J., M. Samaranayake and S. Pradhan (2009). "Epigenetic mechanisms in mammals." Cellular and molecular life sciences **66**(4): 596-612.
Klee, C., T. Crouch and P. Richman (1980). "Calmodulin." Annual review of biochemistry **49**(1): 489-515.
Koch, C. M., R. M. Andrews, P. Flicek, S. C. Dillon, U. Karaöz, G. K. Clelland, S. Wilcox, D. M. Beare, J. C. Fowler and P. Couttet (2007). "The landscape of histone modifications across 1% of the human genome in five human cell lines." Genome research **17**(6): 691-707.
Kolasinska-Zwierz, P., T. Down, I. Latorre, T. Liu, X. S. Liu and J. Ahringer (2009). "Differential chromatin marking of introns and expressed exons by H3K36me3." Nat Genet **41**(3): 376-381.
Koo, A. J., M. Fulda, J. Browse and J. B. Ohlrogge (2005). "Identification of a plastid acyl‐acyl carrier protein synthetase in Arabidopsis and its role in the activation and elongation of exogenous fatty acids." The Plant Journal **44**(4): 620-632.
Kornblihtt, A. R., I. E. Schor, M. Allo and B. J. Blencowe (2009). "When chromatin meets splicing." Nature structural & molecular biology **16**(9): 902-903.
Kouzarides, T. (2007). "Chromatin modifications and their function." Cell **128**(4): 693-705.
Krogan, N. J., J. Dover, S. Khorrami, J. F. Greenblatt, J. Schneider, M. Johnston and A. Shilatifard (2002). "COMPASS, a histone H3 (Lysine 4) methyltransferase required for telomeric silencing of gene expression." Journal of biological chemistry **277**(13): 10753-10755.
Krueger, F., B. Kreck, A. Franke and S. R. Andrews (2012). "DNA methylome analysis using short bisulfite sequencing data." Nature methods **9**(2): 145-151.
Lehnertz, B., Y. Ueda, A. A. Derijck, U. Braunschweig, L. Perez-Burgos, S. Kubicek, T. Chen, E. Li, T. Jenuwein and A. H. Peters (2003). "< i> Suv39h</i>-Mediated Histone H3 Lysine 9 Methylation Directs DNA Methylation to Major Satellite Repeats at Pericentric Heterochromatin." Current Biology **13**(14): 1192-1200.

Li, E., T. H. Bestor and R. Jaenisch (1992). "Targeted mutation of the DNA methyltransferase gene results in embryonic lethality." Cell **69**(6): 915-926.
Lippman, Z., A.-V. Gendrel, M. Black, M. W. Vaughn, N. Dedhia, W. R. McCombie, K. Lavine, V. Mittal, B. May and K. D. Kasschau (2004). "Role of transposable elements in heterochromatin and epigenetic control." Nature **430**(6998): 471-476.
Lister, R., R. C. O'Malley, J. Tonti-Filippini, B. D. Gregory, C. C. Berry, A. H. Millar and J. R. Ecker (2008). "Highly Integrated Single-Base Resolution Maps of the Epigenome in< i> Arabidopsis</i>." Cell **133**(3): 523-536.
Lister, R., M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, J. R. Nery, L. Lee, Z. Ye and Q.-M. Ngo (2009). "Human DNA methylomes at base resolution show widespread epigenomic differences." nature **462**(7271): 315-322.
Long, S. (1991). "Modification of the response of photosynthetic productivity to rising temperature by atmospheric CO2 concentrations: has its importance been underestimated?" Plant, Cell & Environment **14**(8): 729-739.
Luco, R. F., Q. Pan, K. Tominaga, B. J. Blencowe, O. M. Pereira-Smith and T. Misteli (2010). "Regulation of alternative splicing by histone modifications." Science **327**(5968): 996-1000.
Maunakea, A. K., I. Chepelev and K. Zhao (2010). "Epigenome mapping in normal and disease States." Circ Res **107**(3): 327-339.
May, P., W. Liao, Y. Wu, B. Shuai, W. R. McCombie, M. Q. Zhang and Q. A. Liu (2013). "The effects of carbon dioxide and temperature on microRNA expression in Arabidopsis development." Nature communications **4**.
McGhee, J. D. and G. D. Ginder (1979). "Specific DNA methylation sites in the vicinity of the chicken β-globin genes."
Mette, M., W. Aufsatz, J. Van der Winden, M. Matzke and A. Matzke (2000). "Transcriptional silencing and promoter methylation triggered by double‐stranded RNA." The EMBO Journal **19**(19): 5194-5201.
Morgan, D. K. and E. Whitelaw (2008). "The case for transgenerational epigenetic inheritance in humans." Mamm Genome **19**(6): 394-397.
Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer and B. Wold (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq." Nature methods **5**(7): 621-628.
Nathan, D., K. Ingvarsdottir, D. E. Sterner, G. R. Bylebyl, M. Dokmanovic, J. A. Dorsey, K. A. Whelan, M. Krsmanovic, W. S. Lane and P. B. Meluh (2006). "Histone sumoylation is a negative regulator in Saccharomyces cerevisiae and shows dynamic interplay with positive-acting histone modifications." Genes & development **20**(8): 966-976.
Nelson, C. J., H. Santos-Rosa and T. Kouzarides (2006). "Proline isomerization of histone H3 regulates lysine methylation and gene expression." Cell **126**(5): 905-916.
Ng, H. H., F. Robert, R. A. Young and K. Struhl (2003). "Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity." Molecular cell **11**(3): 709-719.
Niu, Y., C. Jin, G. Jin, Q. Zhou, X. Lin, C. Tang and Y. Zhang (2011). "Auxin modulates the enhanced development of root hairs in Arabidopsis thaliana (L.) Heynh. under elevated CO2." Plant, cell & environment **34**(8): 1304-1317.
Nowak, S. J. and V. G. Corces (2004). "Phosphorylation of histone H3: a balancing act between chromosome condensation and transcriptional activation." TRENDS in Genetics **20**(4): 214-220.
Oh, S.-H. and D. M. Roberts (1990). "Analysis of the state of posttranslational calmodulin methylation in developing pea plants." Plant physiology **93**(3): 880-887.
Onder, T. T., N. Kara, A. Cherry, A. U. Sinha, N. Zhu, K. M. Bernt, P. Cahan, B. O. Mancarci, J. Unternaehrer and P. B. Gupta (2012). "Chromatin-modifying enzymes as modulators of reprogramming." Nature **483**(7391): 598-602.

Pavet, V., C. Quintero, N. M. Cecchini, A. L. Rosa and M. E. Alvarez (2006). "Arabidopsis displays centromeric DNA hypomethylation and cytological alterations of heterochromatin upon attack by Pseudomonas syringae." Molecular plant-microbe interactions **19**(6): 577-587.
Raisner, R. M., P. D. Hartley, M. D. Meneghini, M. Z. Bao, C. L. Liu, S. L. Schreiber, O. J. Rando and H. D. Madhani (2005). "Histone variant H2A. Z marks the 5′ ends of both active and inactive genes in euchromatin." Cell **123**(2): 233-248.
Reynolds, L. P. and J. S. Caton (2012). "Role of the pre- and post-natal environment in developmental programming of health and productivity." Mol Cell Endocrinol **354**(1-2): 54-59.
Rhee, I., K. E. Bachman, B. H. Park, K.-W. Jair, R.-W. C. Yen, K. E. Schuebel, H. Cui, A. P. Feinberg, C. Lengauer and K. W. Kinzler (2002). "DNMT1 and DNMT3b cooperate to silence genes in human cancer cells." Nature **416**(6880): 552-556.
Rhee, S. Y., W. Beavis, T. Z. Berardini, G. Chen, D. Dixon, A. Doyle, M. Garcia-Hernandez, E. Huala, G. Lander and M. Montoya (2003). "The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community." Nucleic acids research **31**(1): 224-228.
Riggs, A. D. (1975). "X inactivation, differentiation, and DNA methylation." Cytogenetic and Genome Research **14**(1): 9-25.
Robyr, D., Y. Suka, I. Xenarios, S. K. Kurdistani, A. Wang, N. Suka and M. Grunstein (2002). "Microarray deacetylation maps determine genome-wide functions for yeast histone deacetylases." Cell **109**(4): 437-446.
Ronemus, M. J., M. Galbiati, C. Ticknor, J. Chen and S. L. Dellaporta (1996). "Demethylation-induced developmental pleiotropy in Arabidopsis." Science **273**(5275): 654-657.
Schmitz, R. J. and X. Zhang (2011). "High-throughput approaches for plant epigenomic studies." Current opinion in plant biology **14**(2): 130-136.
Schones, D. E., K. Cui, S. Cuddapah, T. Y. Roh, A. Barski, Z. Wang, G. Wei and K. Zhao (2008). "Dynamic regulation of nucleosome positioning in the human genome." Cell **132**(5): 887-898.
Schones, D. E. and K. Zhao (2008). "Genome-wide approaches to studying chromatin modifications." Nature Reviews Genetics **9**(3): 179-191.
Schübeler, D., D. M. MacAlpine, D. Scalzo, C. Wirbelauer, C. Kooperberg, F. van Leeuwen, D. E. Gottschling, L. P. O'Neill, B. M. Turner and J. Delrow (2004). "The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote." Genes & development **18**(11): 1263-1271.
Shilatifard, A. (2006). "Chromatin modifications by methylation and ubiquitination: implications in the regulation of gene expression." Annu. Rev. Biochem. **75**: 243-269.
Shkolnik‐Inbar, D., G. Adler and D. Bar‐Zvi (2013). "ABI4 downregulates expression of the sodium transporter HKT1; 1 in Arabidopsis roots and affects salt tolerance." The Plant Journal **73**(6): 993-1005.
Shu, W., H. Chen, X. Bo and S. Wang (2011). "Genome-wide analysis of the relationships between DNaseI HS, histone modifications and gene expression reveals distinct modes of chromatin domains." Nucleic acids research **39**(17): 7428-7443.
Smith, A. D., W.-Y. Chung, E. Hodges, J. Kendall, G. Hannon, J. Hicks, Z. Xuan and M. Q. Zhang (2009). "Updates to the RMAP short-read mapping software." Bioinformatics **25**(21): 2841-2842.
Smith, A. D., Z. Xuan and M. Q. Zhang (2008). "Using quality scores and longer reads improves accuracy of Solexa read mapping." BMC bioinformatics **9**(1): 128.
Stedman, E. and E. Stedman (1950). "Cell specificity of histones."
Steger, D. J., M. I. Lefterova, L. Ying, A. J. Stonestrom, M. Schupp, D. Zhuo, A. L. Vakoc, J.-E. Kim, J. Chen and M. A. Lazar (2008). "DOT1L/KMT4 recruitment and H3K79 methylation are ubiquitously coupled with gene transcription in mammalian cells." Molecular and cellular biology **28**(8): 2825-2839.

Sterner, D. E. and S. L. Berger (2000). "Acetylation of histones and transcription-related factors." Microbiology and Molecular Biology Reviews **64**(2): 435-459.
Suter, C. M., D. I. Martin and R. L. Ward (2004). "Germline epimutation of MLH1 in individuals with multiple cancers." Nat Genet **36**(5): 497-501.
Sutherland, J. E. and M. Costa (2003). "Epigenetics and the environment." Ann N Y Acad Sci **983**: 151-160.
Suzuki, M. M. and A. Bird (2008). "DNA methylation landscapes: provocative insights from epigenomics." Nature Reviews Genetics **9**(6): 465-476.
Teng, N., B. Jin, Q. Wang, H. Hao, R. Ceulemans, T. Kuang and J. Lin (2009). "No detectable maternal effects of elevated CO2 on Arabidopsis thaliana over 15 generations." PloS one **4**(6): e6035.
Thurman, R. E., E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano, E. Haugen, N. C. Sheffield, A. B. Stergachis, H. Wang and B. Vernot (2012). "The accessible chromatin landscape of the human genome." Nature **489**(7414): 75-82.
Tompa, R., C. M. McCallum, J. Delrow, J. G. Henikoff, B. van Steensel and S. Henikoff (2002). "Genome-wide profiling of DNA methylation reveals transposon targets of CHROMOMETHYLASE3." Current Biology **12**(1): 65-68.
Trapnell, C., L. Pachter and S. L. Salzberg (2009). "TopHat: discovering splice junctions with RNA-Seq." Bioinformatics **25**(9): 1105-1111.
Trapnell, C., A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn and L. Pachter (2012). "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks." Nature protocols **7**(3): 562-578.
Trapnell, C., B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold and L. Pachter (2010). "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation." Nature biotechnology **28**(5): 511-515.
Turner, B. M. (2000). "Histone acetylation and an epigenetic code." Bioessays **22**(9): 836-845.
Urnov, F. D. and A. P. Wolffe (2001). "Above and within the genome: epigenetics past and present." Journal of mammary gland biology and neoplasia **6**(2): 153-167.
Verhoeven, K. J., J. J. Jansen, P. J. van Dijk and A. Biere (2010). "Stress‐induced DNA methylation changes and their heritability in asexual dandelions." New Phytologist **185**(4): 1108-1118.
Wang, Y., J. Wysocka, J. Sayegh, Y.-H. Lee, J. R. Perlin, L. Leonelli, L. S. Sonbuchner, C. H. McDonald, R. G. Cook and Y. Dou (2004). "Human PAD4 regulates histone arginine methylation levels via demethylimination." Science **306**(5694): 279-283.
Wang, Z., C. Zang, K. Cui, D. E. Schones, A. Barski, W. Peng and K. Zhao (2009). "Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes." Cell **138**(5): 1019-1031.
Wang, Z., C. Zang, J. A. Rosenfeld, D. E. Schones, A. Barski, S. Cuddapah, K. Cui, T. Y. Roh, W. Peng, M. Q. Zhang and K. Zhao (2008). "Combinatorial patterns of histone acetylations and methylations in the human genome." Nat Genet **40**(7): 897-903.
Wolffe, A. P. and J. J. Hayes (1999). "Chromatin disruption and modification." Nucleic acids research **27**(3): 711-720.
Wolffe, A. P. and M. A. Matzke (1999). "Epigenetics: regulation through repression." science **286**(5439): 481-486.
Xu, X., S. Hoang, M. W. Mayo and S. Bekiranov (2010). "Application of machine learning methods to histone methylation ChIP-Seq data reveals H4R3me2 globally represses gene expression." BMC bioinformatics **11**(1): 396.
Yamniuk, A. P. and H. J. Vogel (2004). "Calmodulin's flexibility allows for promiscuity in its interactions with target proteins and peptides." Molecular biotechnology **27**(1): 33-57.
Zhang, X., J. Yazaki, A. Sundaresan, S. Cokus, S. W.-L. Chan, H. Chen, I. R. Henderson, P. Shinn, M. Pellegrini and S. E. Jacobsen (2006). "Genome-wide High-Resolution Mapping and Functional Analysis of DNA Methylation in< i> Arabidopsis</i>." Cell **126**(6): 1189-1201.

Zhang, Y. and D. Reinberg (2001). "Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails." Genes & development **15**(18): 2343-2360.
Zilberman, D., M. Gehring, R. K. Tran, T. Ballinger and S. Henikoff (2007). "Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription." Nature genetics **39**(1): 61-69.

Appendix

5'UTR					
Chromosome	Start	End	Gene loci	Strand	Sample
chr2	2538965	2539093	Parent=AT2G06420.1	+	810ppm
chr4	9560768	9560889	Parent=AT4G16990.5	+	810ppm
genebody					
Chromosome	Start	End	Gene loci	Strand	Sample
chr1	12809077	12809177	AT1G35030	+	430ppm
chr2	600636	600791	AT2G02280	-	430ppm
chr2	6759537	6759587	AT2G15480	+	430ppm
chr3	5862285	5862364	AT3G17185	+	430ppm
chr5	8241535	8241632	AT5G24270	-	430ppm
chr1	15433951	15434118	AT1G40390	-	810ppm
chr1	19741876	19741971	AT1G52990	-	810ppm
chr1	23510087	23510137	AT1G63410	-	810ppm
chr1	9903542	9903749	AT1G28304	+	810ppm
chr1	9903967	9904090	AT1G28304	+	810ppm
chr1	15083848	15083949	AT1G40104	+	810ppm
chr1	15085201	15085305	AT1G40104	+	810ppm
chr1	16239109	16239209	AT1G43145	+	810ppm
chr1	16565906	16566032	AT1G43780	+	810ppm
chr1	18192411	18192618	AT1G49190	+	810ppm
chr1	20311956	20312106	AT1G54420	+	810ppm
chr2	2909423	2909484	AT2G07020	-	810ppm
chr2	5289866	5289977	AT2G12875	-	810ppm
chr2	6555333	6555469	AT2G15110	-	810ppm
chr2	8076354	8076428	AT2G18610	-	810ppm
chr2	14062842	14062976	AT2G33175	-	810ppm
chr2	2020	2089	AT2G01008	+	810ppm
chr2	3616012	3616108	AT2G07981	+	810ppm
chr2	3622575	3622625	AT2G08986	+	810ppm
chr2	6843620	6843722	AT2G15710	+	810ppm
chr3	10255052	10255154	AT3G27680	-	810ppm
chr3	11083082	11083238	AT3G29110	-	810ppm
chr3	15784471	15784604	AT3G43990	-	810ppm
chr3	5867382	5867475	AT3G17190	+	810ppm
chr3	11474161	11474211	AT3G29638	+	810ppm
chr3	11931382	11931489	AT3G30320	+	810ppm
chr3	14195934	14196037	AT3G41761	+	810ppm

chr3	18758146	18758318	AT3G50540	+	810ppm
chr4	10653055	10653123	AT4G19530	+	810ppm
chr5	2129124	2129174	AT5G06850	-	810ppm
chr5	10208372	10208481	AT5G28237	-	810ppm
chr5	15873632	15873690	AT5G39645	-	810ppm
chr5	17231899	17232178	AT5G42955	-	810ppm
chr5	18475670	18475781	AT5G45573	-	810ppm
chr5	3253037	3253167	AT5G10340	+	810ppm
3'UTR					
Chromosome	Start	End	Gene loci	Strand	Sample
chr2	1167442	1167653	Parent=AT2G03821.1	-	810ppm
chr2	9048081	9048220	Parent=AT2G21100.1	-	810ppm
chr2	2503	2591	Parent=AT2G01008.1	+	810ppm
1kb promoter					
Chromosome	Start	End	Gene loci	Strand	Sample
chr2	6483093	6483200	Parent=AT2G15000.5	-	430ppm
chr1	10949903	10950013	Parent=AT1G30820.1	-	810ppm
chr1	17171783	17171851	Parent=AT1G45230.2	-	810ppm
chr1	17171783	17171851	Parent=AT1G45230.1	-	810ppm
chr1	10814321	10814487	Parent=AT1G30530.1	+	810ppm
chr1	11930635	11930807	Parent=AT1G32928.1	+	810ppm
chr1	15890721	15890881	Parent=AT1G42430.2	+	810ppm
chr1	15890721	15890881	Parent=AT1G42430.1	+	810ppm
chr1	20412422	20412615	Parent=AT1G54680.3	+	810ppm
chr1	20412422	20412615	Parent=AT1G54680.2	+	810ppm
chr1	22563481	22563591	Parent=AT1G61210.2	+	810ppm
chr2	184025	184294	Parent=AT2G01420.2	-	810ppm
chr2	455680	455778	Parent=AT2G01970.1	-	810ppm
chr2	1167442	1167653	Parent=AT2G03820.1	-	810ppm
chr2	4661488	4661646	Parent=AT2G11623.2	-	810ppm
chr2	9048081	9048220	Parent=AT2G21090.1	-	810ppm
chr2	10317854	10318001	Parent=AT2G24255.1	-	810ppm
chr2	11076089	11076173	Parent=AT2G25970.1	-	810ppm
chr2	17418019	17418119	Parent=AT2G41740.1	-	810ppm
chr2	1258469	1258533	Parent=AT2G03980.1	+	810ppm
chr2	7295013	7295151	Parent=AT2G16835.1	+	810ppm
chr2	9262804	9262854	Parent=AT2G21655.1	+	810ppm
chr2	9470659	9470822	Parent=AT2G22300.2	+	810ppm

chr2	9470659	9470822	Parent=AT2G22300.1	+	810ppm
chr2	9791207	9791257	Parent=AT2G23000.1	+	810ppm
chr2	18498534	18498652	Parent=AT2G44850.2	+	810ppm
chr3	48457	48507	Parent=AT3G01140.1	-	810ppm
chr3	10262338	10262461	Parent=AT3G27700.2	-	810ppm
chr3	11010199	11010386	Parent=AT3G29020.2	-	810ppm
chr3	13494376	13494508	Parent=AT3G32940.1	-	810ppm
chr3	15245856	15245996	Parent=AT3G43300.2	-	810ppm
chr3	20127277	20127391	Parent=AT3G54350.3	-	810ppm
chr3	20127277	20127391	Parent=AT3G54350.2	-	810ppm
chr3	20353431	20353613	Parent=AT3G54930.1	-	810ppm
chr3	8574593	8574688	Parent=AT3G23790.1	+	810ppm
chr3	10464455	10464554	Parent=AT3G28130.2	+	810ppm
chr3	10827178	10827309	Parent=AT3G28820.1	+	810ppm
chr3	11727201	11727330	Parent=AT3G29810.1	+	810ppm
chr3	18172647	18172697	Parent=AT3G49030.2	+	810ppm
chr4	5844726	5844809	Parent=AT4G09170.1	-	810ppm
chr4	7100460	7100589	Parent=AT4G11800.1	-	810ppm
chr4	7322664	7322847	Parent=AT4G12340.1	-	810ppm
chr4	1091636	1091748	Parent=AT4G02485.1	+	810ppm
chr4	1373198	1373305	Parent=AT4G03100.1	+	810ppm
chr4	2445133	2445222	Parent=AT4G04830.2	+	810ppm
chr4	2445133	2445222	Parent=AT4G04830.1	+	810ppm
chr4	5238035	5238177	Parent=AT4G08290.1	+	810ppm
chr4	5238035	5238177	Parent=AT4G08290.2	+	810ppm
chr4	6288354	6288446	Parent=AT4G10060.1	+	810ppm
chr4	6617736	6617857	Parent=AT4G10750.1	+	810ppm
chr4	9560768	9560889	Parent=AT4G16990.5	+	810ppm
chr5	12566170	12566280	Parent=AT5G33300.1	-	810ppm
chr5	16030542	16030627	Parent=AT5G40040.1	-	810ppm
chr5	21422220	21422311	Parent=AT5G52860.1	-	810ppm
chr5	22186885	22186989	Parent=AT5G54610.1	-	810ppm
chr5	24002102	24002241	Parent=AT5G59560.1	-	810ppm
chr5	24002102	24002241	Parent=AT5G59560.2	-	810ppm
chr5	24002872	24003017	Parent=AT5G59560.1	-	810ppm
chr5	24002872	24003017	Parent=AT5G59560.2	-	810ppm
chr5	1518869	1518944	Parent=AT5G05140.1	+	810ppm
chr5	5352266	5352318	Parent=AT5G16350.1	+	810ppm
chr5	6789369	6789472	Parent=AT5G20100.1	+	810ppm
chr5	8272817	8272908	Parent=AT5G24310.1	+	810ppm
chr5	8272817	8272908	Parent=AT5G24310.2	+	810ppm

chr5	11379476	11379583	Parent=AT5G30495.2	+	810ppm
chr5	11379476	11379583	Parent=AT5G30495.1	+	810ppm
chr5	13897210	13897307	Parent=AT5G35732.1	+	810ppm
chr5	17190860	17191088	Parent=AT5G42880.1	+	810ppm
chr5	20640243	20640300	Parent=AT5G50750.1	+	810ppm

Table S 1 The gene elements (5'UTR, genebody, 3'UTR and 1kb promoter) overlapped with DMRs. Start and end are the geneome coordinates of DMRs. Sample indicate the strain which has higher DNA methylation level.

Marks subgroup	First scan	Second scan	Third scan	Forth scan	Fifth scan
Promoter	0.873	0.887	0.894	0.899	0.902
Structural	0.862	0.875	0.883	0.886	0.888
Repressive	0.507	0.573	0.609	0.624	0.632
Distal/other	0.675	0.727	0.762	0.779	0.79
Promoter + Structural	0.879	0.894	0.9	0.904	0.907
Promoter + Repressive	0.874	0.887	0.894	0.9	0.903
Promoter + Distal	0.878	0.892	0.897	0.902	0.906
All HM	0.883	0.896	0.902	0.907	0.91
All HM + DNaseI	0.891	0.906	0.913	0.917	0.918

Table S 2 Pearson's r for different combinations of chromatin features in each
The Pearson's r values were averaged by 10-fold cross-validation. Promoter m
H3K4me2, H3K4me3,H2A.Z, H3K9ac and H3K27ac ; structural marks:
and H3K79me2; repressive marks: H3K27me3 and H3K9me3and distal/
H3K4me1, H4K20me1 and H3K9m (Raisner et al., 2005; Barski et al., 2
Benevolenskaya 2007; Koch et al., 2007; Steger et al., 2008; Kolasinska
2009).

	Coefficients
(Intercept)	-0.806
h(4.93013-Dnase.1)	-0.658
h(H3k4me1.1-3.67525)	-0.627
h(3.67525-H3k4me1.1)	1.829
h(H3k4me2.2-3.86258)	-0.213
h(3.86258-H3k4me2.2)	-0.375
h(H3k79me2.2-3.86846)	0.515
h(3.86846-H3k79me2.2)	-0.261
h(H3k9ac.4-1.51068)	0.859
h(1.51068-H3k9ac.4)	0.894
h(H3k4me1.5-0.524592)	0.158
h(0.524592-H3k4me1.5)	-0.077
h(H3k79me2.5-2.81558)	-0.186
h(2.81558-H3k79me2.5)	0.048
h(Dnase.1-4.93013) * h(H3k4me2.1-5.04162)	0.703
h(Dnase.1-4.93013) * h(5.04162-H3k4me2.1)	-0.788
h(4.93013-Dnase.1) * h(H3k9ac.1-4.16879)	-0.415
h(Dnase.1-0.23396) * h(3.86846-H3k79me2.2)	-0.069
h(0.23396-Dnase.1) * h(3.86846-H3k79me2.2)	0.153
h(4.93013-Dnase.1) * h(Dnase.4-2.0571)	-0.230
h(Dnase.1-2.83178) * h(H3k9ac.4-1.51068)	-0.141
h(2.83178-Dnase.1) * h(H3k9ac.4-1.51068)	-0.053
h(H2az.1-4.85662) * h(H3k9ac.4-1.51068)	-0.197
h(4.85662-H2az.1) * h(H3k9ac.4-1.51068)	-0.095

h(3.67525-H3k4me1.1) * h(H3k4me2.1-1.80456)	0.456
h(3.67525-H3k4me1.1) * h(1.80456-H3k4me2.1)	-0.368
h(3.67525-H3k4me1.1) * h(H2az.4-3.41056)	-0.574
h(3.67525-H3k4me1.1) * h(3.41056-H2az.4)	0.233
h(3.67525-H3k4me1.1) * h(H3k9ac.4-2.85807)	-0.200
h(3.67525-H3k4me1.1) * h(2.85807-H3k9ac.4)	-1.092
h(H3k36me3.3-1.17832) * h(H3k9ac.4-1.51068)	0.165
h(1.17832-H3k36me3.3) * h(H3k9ac.4-1.51068)	-0.767
h(H3k9me3.3-0.800334) * h(H3k9ac.4-1.51068)	-0.319
h(0.800334-H3k9me3.3) * h(H3k9ac.4-1.51068)	0.106

Table S 3 Features selected in the final model. The number ".1" to ".5" indicate the bin#21, bin#28, bin#41, bin#24 and bin#13 selected by scanning step respectively.