DATA MINING TO IDENTIFY GENE REGULATORY ELEMENTS

by

LEXIANG JI

(Under the Direction of Ping Ma)

ABSTRACT

Gene regulatory elements are essential for the survival and development of all organisms but only a handful of gene regulatory elements have been discovered thus far in plants. Unfortunately, traditional methods to identify regulatory elements are ineffective as these regions are typically short (4-12 base pairs) and can be thousands of base pairs away from target genes. My project aimed to identify regulatory elements in plant genomes by mining a large and diverse set of histone modifications, chromatin modifications, and chromatin accessibility data with statistical approaches including K-means clustering and smoothing spline clustering. The results revealed an abundance of distal accessible chromatin regions in the maize genome which contain distinct combinations of chromatin modifications. Results from this project provide valuable clues for the future improvement of economically important crop traits.

INDEX WORDS: Regulatory elements; Chromatin; Statistics; K-means clustering; Smoothing spline clustering

DATA MINING TO IDENTIFY GENE REGULATORY ELEMENTS

by

LEXIANG JI

BS, Zhejiang A & F University, China, 2010

MS, Beijing Forestry University, China, 2013

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2019

DATA MINING TO IDENTIFY GENE REGULATORY ELEMENTS

by

LEXIANG JI

Major Professor:	Ping Ma
Committee:	Wenxuan Zhong
	Robert J. Schmitz

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2019

ACKNOWLEDGEMENTS

I'm very grateful to have Ping as my thesis advisor. He is not only a preeminent scientist but also an outstanding advisor on many fronts. I have benefited a lot from the process of learning statistics with him which dramatically changed my understanding of how to use statistical methods to tackle scientific questions. His patience and guidance lead to the completion of my thesis.

I would also like to thank other two committee members, Wenxuan and Bob. They gave lots comments and suggestions during my thesis research and provided numerous advice for my career development. In addition, Bob is the keystone of my entire graduate career. He is the person who provided me a systematical training in science and granted me the chance to improve my ability in statistics. Moreover, I would like to thank Bill, Zefu, and Tina for providing tissues and preparing libraries. Finally, I would also like to thank all members from Ma's lab for their continued support during my statistical studies. It's been a great journey.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

BACKGROUND AND LICTURE REVIEW

A fundamental question in biology is the understanding of how gene expression is regulated. At the center of this question is the identification and characterization of gene regulatory elements in the genome. Unfortunately, methods to experimentally identify regulatory elements individually are slow, expensive, labor intensive and therefore not scalable (Studer et al. 2011). Traditional computation methods to predict regulatory elements based on sequence information alone are ineffective as regulatory elements are typically short (4-12 base pairs) and can be hundreds of thousands of bases pairs away from their target genes.

My project aimed to identify regulatory elements by mining a large and diverse set of chromatin structure and modification data, including histone modifications, chromatin modifications, chromatin accessibility and transcriptomes. Clustering methods were adopted to identify accessible chromatin regions with different combinations of histone modifications. In this project, data were generated from four different sequencing technologies including i) **ATAC-seq** (assay for transposase-accessible chromatin using sequencing), ii) **MethylC-seq** (whole-genome bisulfite sequencing), iii) **RNA-seq** (whole transcriptome shotgun sequencing), and iv) **ChIP-seq** (chromatin immunoprecipitation with sequencing).

Chromatin accessibility and ATAC-seq. In eukaryotic genomes, DNA wraps around nucleosomes which forms the basic unit of chromatin (Richmond and Davey 2003). The dynamic distribution of nucleosomes results in different levels of chromatin accessibility, which influences the ability of transcription factors to bind *cis*-elements (Kaplan et al. 2009). Thus, genome-wide identification of accessible regions forms a foundation of information to predict potential regulatory elements bound by transcription factors. Currently, the most advanced technology to determine the chromatin accessibility in a genome-wide scale is ATAC-seq (Buenrostro et al. 2013). This method has recently been adopted for studies of plant genomes (Lu et al. 2017).

DNA methylation and MethylC-seq. DNA methylation is a key chromatin modification and has been found in all studied plants (Du et al. 2015; Niederhuth et al. 2016). It is associated with the silencing of transposable elements (TEs), repeats and certain genes (Law and Jacobsen 2010). In plants, DNA methylation can be separated into three sequence contexts: CG, CHG, and CHH (H = A, T, and C). These different contexts are maintained by different enzymatic pathways (Law and Jacobsen 2010; Stroud et al. 2013; Matzke and Mosher 2014; Du et al. 2015). In mammalian genomes, low-methylated or unmethylated regions are often potential regulatory regions (Jones 2012). Thus, the levels of methylation may be indicative of the regulatory element potential of a locus.

MethylC-seq is a whole genome bisulfite sequencing technology and regarded as the gold standard to detect DNA methylation at single-base resolution. This method was initially applied to the modern plant, *Arabidopsis thaliana* (Cokus et al. 2008; Lister et al.

2008). Using this method, unmethylated cytosines will be converted to uracils by bisulfite treatment. Finally, after polymerase chain reaction (PCR) amplification and deep sequencing, methylated and unmethylated cytosines will be read out as cytosines and thymines, respectively (Hayatsu et al. 1970; Shapiro et al. 1973; Shapiro et al. 1974; Hayatsu 1976; Frommer et al. 1992; Clark et al. 1994; Clark et al. 2006; Ji et al. 2014).

Transcriptional profile and RNA-seq. The Central Dogma posits that DNA is transcribed to RNA and RNA is translated to protein. Although the function of most genes is carried out by assembled proteins, we often use the abundance of RNA to reflect the activity of respected proteins. Due to the limitation of current sequencing technologies and gene annotation methods, there are expressed genes with coding capabilities yet have not been fully annotated in the intergenic regions. Additionally, actively transcribed loci encoding long non-coding RNAs may also hamper the identification of potential regulatory elements as they also are characterized by accessible chromatin. Therefore, transcriptome data can be used to help distinguish potential regulatory elements from transcribed protein-coding genes and long non-coding RNAs. Many different methods have been developed to quantify the transcription levels of target regions such as quantitative real-time PCR (qRT-PCR) or whole transcriptome sequencing such as microarray or RNA-seq (Malone and Oliver 2011). With the decrease of sequencing cost in the last ten years, RNA-seq has become the standard approach in determining the expression levels in a genome-wide scale.

Histone modifications and ChIP-seq. In eukaryotes, a single nucleosome is comprised of an octamer of four histones, including H2A, H2B, H3, and H4 (Richmond and Davey 2003). A nucleosome includes a dimer of each histone. Post-translational modifications of tails of histones are involved in the regulation of gene expression in a variety of ways. Many different types of histone modifications have been discovered in mammals and plants, including acetylation, methylation, phosphorylation, and ubiquitylation (Strahl and Allis 2000; Berger 2002). H3K4me3, or tri-methylation of histone H3, is associated with expressed genes and is deposited around the transcription start site (TSS) of activated genes. This modification is absent from transposable elements and repetitive sequences (Zhang et al. 2009). Similar to the function of trimethylated H3K4, H3K36me3 is prevalent in transcribed genes. This modification is often colocalized with H3K4me3 and related to transcriptional initiation and elongation (Wagner and Carpenter 2012). Unlike methylation of H3K4 and H3K36, H3K27me3 has been found in heterochromatin and is associated with transcriptional repression. Thousands of genes are targeted by H3K27me3 in model plant species, *A. thaliana* (Johnson et al. 2004; Zhang et al. 2007a). H3K56 acetylation (H3K56ac) is an additional indicator of active transcription and is enriched around promoter regions in plants (Tanurdzic et al. 2008). Currently, ChIP-seq is the standard method to identify genome-wide distributions of DNA-bound factors and histone modifications (Park 2009). This method relies on antibodies specific for proteins or histones with specific tail modifications.

CHAPTER 2

IDENTIFICATION OF CHROMATIN ACCESSIBLE REGIONS

ATAC-seq data generated from maize leaf tissue was aligned to the maize B73v4 genome and 32,111 accessible chromatin regions (ACRs) were identified (**Figure 1**). The identified ACRs ranges mostly from 200-400 bp and the median length is 275 bp (**Figure 2**), accounting for ~1% of the maize genome. Among them, 12,495 (38.9%) were in or overlapped with genic regions (referred to as genic ACRs (gACRs)). Another 9,183 (28.6%) ACRs were within 2 kilobase pairs (kb) of genic regions (referred to as proximal ACRs (pACRs)). In addition, 10,433 (32.5%) ACRs were at least 2 kb away from genic region (referred to as distal ACRs (dACRs)) (**Figure 3**). These dACRs were much more GC-rich relative to the negative control regions ("Control", non-ACR intergenic regions, see methods) with an average GC content of 51.5% (**Figure 4**). The higher GC content in dACRs can facilitate well-positioned nucleosomes and are less likely to be methylated. Furthermore, consistent with the evaluated accessibility, the DNA cytosine methylation in all sequence contexts was markedly reduced at dACRs (**Figure 5**). Taken together, the dACRs identified were utilized to identify features of distal regulatory regions using statistical methods.

CHAPTER 3

APPLICATION OF CLUSTERING METHODS

GENERATION OF INPUT DATA FOR CLUSTERING

In this project, normalized values of twenty 100-bp bins from upstream and downstream of dACR summits were extracted for the histone modifications H3K27me3, H3K36me3, H3K4me3, and H3K56ac. The normalized values were then concatenated into a single matrix with 160 columns.

In addition, to reduce the computational complexity, 500 dACRs were randomly selected from identified dACRs for subsequent clustering analysis. Briefly, the input can be written as follows

$$\boldsymbol{y}_{ij} = (\ y_{i1}, y_{i2}, y_{i3}, y_{i4})$$

we let the $\boldsymbol{y}_{ij} = (\ y_{ij1}, \ldots, y_{ijn}\ )$ be the j-th modification for the i-th dACR, where $n = 40$, j = (1, 2, 3, 4), representing H3K27me3, H3K36me3, H3K4me3, and H3K56ac, respectively.

K-MEANS CLUSTERING

Algorithm

In statistical data analyses, we often face a question of grouping a set of data points or observations into multiple clusters. To date, numerous clustering algorithms have been developed to address various types of scientific questions (Xu and Tian 2015). K-means clustering method is one such algorithm and has been widely used in different fields (Hartigan and Wong 1979; Jain 2010). The method requires a pre-defined number of clusters and can rapidly find local optimums. The principle of this method is to minimize within-cluster sum of squares over defined $\boldsymbol{K}$ clusters and the computational difficulty is NP-hard.

The method can be written as the following equation:

$$W(C_j) = \sum_{X_i \in C_j} ||X_i - \mu_j||^2$$

Where the $\boldsymbol{X} = \{x_i, i = 1..n\}$, x_i is the i-th value and $\boldsymbol{N}$ is the total number of data points in collected dataset. $\mu = \{\mu_j, i = 1..k\}$, μ_j is the mean of cluster j. $C = \{c_j, i = 1..k\}$, $W(C_j)$ is the within-cluster sum of squares of cluster j.

$$W = \sum_{j=1}^{K} \sum_{X_i \in C_j} ||X_i - \mu_j||^2 \qquad (1)$$

Where W in the model 1 is the total within-cluster sum of squares of clustered $\boldsymbol{K}$ groups. The main steps of the K-means clustering method include:

Algorithm: K-means clustering method

Input: samples $\boldsymbol{X} = \{x_1, \ldots, x_n\}$, number of clusters $\boldsymbol{K}$, maximum number of iterations $\boldsymbol{m}$

Generate $\boldsymbol{K}$ random initial "means" within $\boldsymbol{X} = \{x_1, \ldots, x_n\}$

for i = 1 to $\boldsymbol{m}$ do

Assign each sample to the nearest mean with the least squared Euclidean distance

The centroid of each cluster becomes the new "means"

exit and **return** clustered samples **if** the convergence has been reached

end

Clustering result

Calculation of within-cluster sum of squares. To determine the number of clusters, the total within-cluster sum of squares was computed from different sets of $\boldsymbol{K}$, ranging from 2 to 15 (**Table 1**). The total within-cluster sum of squares dropped significantly when the data was separated from 2 to 5 clusters and leveled out following five clusters. Finally, manual inspection of grouped dACRs and associated histone modifications revealed the $\boldsymbol{K}$ = 5 has the best performance.

Table 1 Minimized within-cluster sum of pre-defined *K* clusters

Predefined $\boldsymbol{K}$	Maximum number of iterations	WCSS
1	/	943.71
2	30	780.10
3	30	644.49
4	30	542.08
5	30	501.42
6	30	472.71
7	30	449.33

8	30	433.29
9	30	418.03
10	30	405.61
11	30	394.12
12	30	385.58
13	30	378.90
14	30	372.16
15	30	365.48

Identification of dACRs with distinct combinations of chromatin modifications. Using K-means clustering approach, five empirical clusters were identified from dACRs in the maize genome, which were resolved into four groups. The majority of dACRs (51.4%) were depleted of histone modifications (Unmodified group) (**Figure 6**), 8.8% of dACRs contained primarily H3K27me3 (K27me3 group) (**Figure 7**), which is known to be associated with transcriptional repression. Another 12.0% of dACRs contained primarily H3K56ac (Kac group) (**Figure 8**), which is an indicator of active transcription. The last 27.8% of dACRs contained a number of histone modifications (Transcribed group) (**Figure 9**), including H3K4me3, H3K36me3, and H3K56ac. These histone modifications are typically found at transcribed genes. In addition, abundant transcripts were produced from flanking regions of the transcribed group which colocalized with these histone modifications (**Figure 9**). These results suggested the transcribed group may be associated with unannotated genes rather than intergenic regulatory elements.

SMOOTHING SPLINE CLUSTERING

Algorithm

Smoothing spine clustering is a clustering method for large-scale functional data with multiple covariates developed by (Ma et al. 2006). The key feature of this clustering method compared to the K-means clustering method is that we assume that the response is a curve rather than a discrete, finite dimensional vector. Thus, the response is referred to as functional data. We assume the functional data of i-th observation $\boldsymbol{y}_i = (y_{i1}, \dots, y_{in_i})$ follows the mixed-effect model as below.

$$\boldsymbol{y}_i = \mu(\boldsymbol{x}_i) + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \qquad (2)$$

The Model 2 is estimated using the popular least squares through the minimization of

$$\sum_{i=1}^{n}\left([\boldsymbol{y}_i - \mu(\boldsymbol{x}_i) - \boldsymbol{Z}_i\boldsymbol{b}_i]^T[\boldsymbol{y}_i - \mu(\boldsymbol{x}_i) - \boldsymbol{Z}_i\boldsymbol{b}_i] + \sigma^2\boldsymbol{b}_i^T\boldsymbol{B}^{-1}\boldsymbol{b}_i^T\right) + N\lambda M(\mu),$$

where $N = \sum_{i=1}^{n} n_i$, the $M(\mu)$ is a quadratic function that quantifies the roughness of μ, λ is the smoothing parameters that control the trade-off between the goodness of fit and the smoothness of μ. The tuning parameter λ will be selected using the popular generalized cross validation (GCV) method. For later purpose, we introduce the smoothing matrix $\boldsymbol{A}$ such that $\widehat{\boldsymbol{y}} = \boldsymbol{A}\boldsymbol{y}$.

In order to incorporate the mixture model for clustering purpose, we modify Model 2 as follows,

$$\boldsymbol{y}_i = \mu_k(\boldsymbol{x}_i) + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \text{ with probability } p_k,$$

where $k = 1, \dots K$ is the index for clusters, cluster probabilities satisfy $\sum_k p_k = 1$, the population mean μ_ks are assumed to be a smooth function on the domain of the predictor

vector, $\boldsymbol{x}_i$, $\boldsymbol{b}_i \sim N(\mathbf{0}, \boldsymbol{B})$. The predictor vector is a $p \times 1$ random-effects vector associated with the $n_i \times p$ design matrix $\boldsymbol{Z}_i$, and random error $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$, which is independent of $\boldsymbol{b}_i$.

In this project, we let $\boldsymbol{y}_i = (y_{i1}, \dots, y_{in_i})$ be the modification for the i-th dACR, and $n_i = 160$. We let $\boldsymbol{x}_i = (\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i})$, where $\boldsymbol{x}_{1i}$ is the vector for location on the genome, and $\boldsymbol{x}_{2i}$ is the indicator vector indicating the type of modification.

In order to accommodate the correlation between $y_{i1}, \dots, y_{in_i}$ and model heterogeneity among $\boldsymbol{y}_1, \dots \boldsymbol{y}_N$ we let $p = 1$, and thus $\boldsymbol{b}_i$ is a scalar, $\boldsymbol{B} = \sigma_b^2$, and $\boldsymbol{Z}_i = \mathbf{1}$. Then we have the same correlation across time. For computation, we write out the mixture Henderson's likelihood

$$\sum_{i=1}^{N} \log \sum_{k=1}^{K} [p_k f_y(y_i; b_i, J_{ik} = 1) f_b(b_i; J_{ik} = 1)],$$

where J_{ik} is the latent cluster labeling variable such that $J_{ik} = 1$ if the i-th peak belongs to the k-th cluster and 0 otherwise. The negative penalized Henderson's likelihood is written as

$$\begin{aligned} constant - \sum_{i=1}^{n} \sum_{k=1}^{K} J_{ik} \log p_k \\ + \frac{1}{2\sigma^2} \sum_{i=1}^{n} \sum_{k=1}^{K} J_{ik} ([\boldsymbol{y}_i - \mu_k(\boldsymbol{x}_i) - \boldsymbol{Z}_i \boldsymbol{b}_i]^T [\boldsymbol{y}_i - \mu_k(\boldsymbol{x}_i) - \boldsymbol{Z}_i \boldsymbol{b}_i] \\ + \sigma^2 \boldsymbol{b}_i^T \boldsymbol{B}_k^{-1} \boldsymbol{b}_i^T) + \sum_{k=1}^{K} N \lambda_k M(\mu_k). \end{aligned}$$

Then the estimation of the parameter and the tuning of smoothing parameters is calculated with the expectation–maximization (EM) algorithm combined with

generalized cross-validation (GCV) as outlined in (Craven and Wahba 1978; Gu and Ma 2005)

The initial cluster means are stimulated multiple times as different chains. For the sake of completeness, we state the criterion for selecting the chain as below. Briefly, we adopt the Bayesian information criterion (BIC) proposed in (Schwarz 1978) as below

$$BIC = -2\sum_{i=1}^{N}\log\sum_{k=1}^{K}\left[p_k\psi_y(\boldsymbol{y}_i;\mu_k(\boldsymbol{x}_i),\boldsymbol{\Sigma}_k)\right] + \left(\sum_{k=1}^{K} tr(\boldsymbol{A}_k) + 4K\right)\log N,$$

where ψ is the pdf for multivariate normal distribution and $\boldsymbol{A}_k$ is the smoothing matrix in cluster k, $\boldsymbol{\Sigma}_k = \boldsymbol{Z}_i B_k \boldsymbol{Z}_i^T + \sigma^2 \boldsymbol{I}$.

The optimal chain is selected so that the BIC is minimized.

Clustering result

Main groups are reproduced from different clustering methods. The K-means clustering method determined that the best performance was achieved when the cluster = 5. Therefore, in order to compare the results between K-means clustering and smoothing spline clustering, the dACRs were separated into five clusters directly using smoothing spline clustering without additional pre-selection.

Four main groups (unmodified, K27me3, Kac, and transcribed) were readily reproduced from smoothing spline clustering. Using smoothing spline clustering method, 29.6% of dACRs resembled the chromatin attributes of unmodified group, 10.4% of dACRs resembled the chromatin attributes of K27me3 group, 36.8% of dACRs resembled the chromatin attributes of Kac group, and 23.2% of dACRs resembled the chromatin attributes of transcribed group.

COMPARSION OF CLUSTERING METHODS

Although the same four main groups were revealed from different clustering methods, it was still unclear whether these groups are composed of the same dACRs. Therefore, a direct comparison of constituted dACRs was performed. In the view of groups identified by K-means clustering method, the majority of dACRs in the K27me3 group (100.0%), Kac group (95.0%), and transcribed groups (83.5%) were classified into the same groups clustered by smoothing spline clustering approach. However, 41.6% of dACRs in the unmodified group were classified into the Kac group, suggesting a relatively high similarity among these two groups. Therefore, future studies may be needed to include additional histone modifications to increase the sensitivity of clustering.

Table 2 Comparison of two clustering methods

		Smoothing spline clustering			
	Group	unmodified	K27me3	Kac	transcribed
K-means clustering	unmodified	57.6%	0.8%	41.6%	0.0%
	K27me3	0.0%	100.0%	0.0%	0.0%
	Kac	0.0%	5.0%	95.0%	0.0%
	transcribed	0.0%	2.2%	14.4%	83.5%

CHAPTER 4

CONCLUSION

An exciting discovery from my project is that potential regulatory elements are associated with certain chromatin structures and are accessible when regulation is required. Results from this study led to the discovery of several types of regulatory elements which may be potential enhancer or repressors. The group with histone modification H3K27me3 may be linked with repression and the group with histone acetylation mark H3K56ac may be associated with activation. These results strongly suggest that specific transcription factor-DNA interactions cooperate with chromatin modifications to regulate gene expression, which has not previously described in plants genome-wide. Further studies are needed to examine these hypotheses.

MATERIALS AND METHODS

Plant material and growth conditions. *Z. mays* L., cultivar B73, was grown from seed collected from field-grown ears grown during the summer of 2017 in Athens, Georgia. ATAC-seq, ChIP-seq, MethylC-seq, and RNA-seq experiments were all performed on seedling tissue grown under the following conditions: kernels were sown in Sungro Horticulture professional growing mix (Sungro Horticulture Canada Ltd., 52130 RR 65, P.O. Box 189, Seba Beach, AB T0E 2B0 Canada). Soil was saturated with tap water and placed under a 50/50 mixture of 4100K (Sylvania Supersaver Cool White Delux F34CWX/SS, 34W) and 3000K (GE Ecolux w / starcoat, F40CX30ECO, 40W) light. The photoperiod was 16 hours of light, eight hours of dark. The temperature was approximately 25°C during light hours. The relative humidity was approximately 54%. Seedlings were grown for approximately six days and harvested from four to six hours after photoperiod dawn. Seedlings were harvested when the first leaf had emerged two-to-three centimeters above the apical tip of the coleoptile. The seedlings were cut three millimeters above the coleoptile-mesocotyl boundary, excluding the shoot apical meristem, and the second leaf was removed from within the sheath of the first leaf. Only the inner second leaves, which contains the third and fourth leaves sheathed inside, were used for experiments.

Library preparation. ATAC-seq library preparation was performed as described in (Lu et al. 2017). ChIP-seq library preparation was performed following the general protocol of (Zhang et al. 2007b). MethylC-seq library preparation was performed as previously described in (Urich et al. 2015). RNA-seq library preparation: second leaves were flash-frozen with liquid N_2 immediately after collection. Samples were ground to a powder with a mortar and pestle in liquid N_2. Total RNA was extracted and purified with TRIzol™ Reagent (Thermo Fisher Scientific) following the manufacturer's protocol. For each tissue and replicate, 1.3 μg of total RNA was prepared for sequencing with the Illumina Truseq mRNA Stranded Library Kit (Illumina, San Diego, CA) following the manufacturer's instructions.

Sequencing Information. Sequencing of ATAC-seq, ChIP-seq, and RNA-seq were performed at the University of Georgia Genomics Facility using an Illumina NextSeq 500 instrument. MethylC-seq was performed at the University of Minnesota, Twin Cities using an Illumina HiSeq 2500 instrument. ATAC-seq and MethylC-seq were sequenced in paired-end 35 bp and 125 bp, respectively. ChIP-seq was sequenced in single end 75 bp.

ATAC-seq raw data processing and alignment. Raw reads were trimmed with Trimmomatic v0.33 (Bolger et al. 2014). Reads were trimmed for NexteraPE with a maximum of two seed mismatches, palindrome clip threshold of 30, and simple clip threshold of 10. Reads shorter than 30 bp were discarded. Trimmed reads were aligned to the *Z. mays* AGPv4 reference genome using Bowtie 1.1.1 with the following parameters:

"bowtie -X 1000 -m 1 -v 2 --best –strata"(Langmead et al. 2009). Aligned reads were sorted using SAMtools v1.3.1 (Li et al. 2009) and clonal duplicates were removed using picard (http://broadinstitute.github.io/picard).

RNA-seq raw data processing, alignment, and expression quantification. Raw reads were trimmed with Trimmomatic v0.33 with default parameters (Bolger et al. 2014). Qualified reads were aligned to the the *Z. mays* AGPv4 reference genome using hisat2 v2.0.5 (Kim et al. 2015). Gene expression values were computed using StringTie v1.3.3b with the annotation version AGPv4.38 (Pertea et al. 2015).

MethylC-seq raw data processing, alignment, and calculation of methylation status. Quality-filtering and adapter-trimming were performed using cutadapt (Martin 2011). Reads were aligned to the *Z. mays* AGPv4 reference genome using Methylpy 1.3 as previously described in (Schultz et al. 2015). Chloroplast DNA (which is fully unmethylated) was used as a control to calculate the sodium bisulfite reaction non-conversion rate of unmodified cytosines. The conversion rates were > 99.9%. A binomial test was used to determine the methylation status of cytosines with a minimum coverage of three reads.

ChIP-seq raw data processing and alignment. Raw reads were trimmed with Trimmomatic v0.33 with default parameters (Bolger et al. 2014). Qualified reads were aligned to the *Z. mays* AGPv4 reference genome using Bowtie 1.1.1 with following parameters: "bowtie -m 1 -v 2 --best --strata --chunkmbs 1024 -S" (Langmead et al.

2009). Aligned reads were sorted using SAMtools v1.2 and clonal duplicates were removed using SAMtools v0.1.19 (Li et al. 2009).

Heatmap and metaplot analysis. Briefly, 200 10-bp bins were created for both upstream and downstream regions of identified dACRs. For MethylC-seq, weighted methylation levels were computed for each predetermined bin (Schultz et al. 2012). For ChIP-seq and RNA-seq samples, the number of reads per bin were normalized by total aligned reads in each library. Average values were calculated for samples with replicates. Histone modifications were further normalized by subtracting H3 from the values. Normalized values lower than zero were set to 0. Finally, the 95th quantile value of each sample was set as an upper limit. The average values of each bin were used to construct metaplots.

Definition of intergenic negative control regions. To create the intergenic negative control regions, we first identified the uniquely mappable regions by re-mapping all possible simulated 75 bp fragments from the *Z. mays* v4 AGPv4 reference genome (Jiao et al. 2017), with the same parameters for ChIP-seq analysis. Genomic regions with mapped reads were considered as uniquely mappable. Annotated genes and their 2 kb nearby regions, as well as gene-distal ACRs, were removed. Negative control regions with the same length distribution to dACRs were then generated by the "shuffle" command in BEDTools (Quinlan and Hall 2010).

Identification of accessible chromatin regions (ACRs). MACS2 (Zhang et al. 2008) was used to define accessible chromatin regions (ACRs) with the "--keep-dup all" function. To find high quality ACRs, the following filtering steps were performed: Generally, 1) ACRs called with MACS2 were split into 50 bp windows with 25 bp steps; 2) the Tn5 integration frequency in each window was calculated and normalized with the average frequency in the total genome; 3) windows passing the integration frequency cutoff (25-fold) were merged together with 150 bp gaps; 4) small regions with only one window were then filtered with "length >50bp". The sites within ACRs with the highest Tn5 integration frequency were defined as "summits".

Definition of gACRs, pACRs, and dACRs. ACRs were separated into three groups based on their proximity relative to known annotated genes in *Z. mays* AGPv4.38 genome annotation (Jiao et al. 2017). Proximal regions were within 2 kb up- or downstream of annotated genes but no direct overlap. Distal regions were greater than 2 kb from annotated genes.

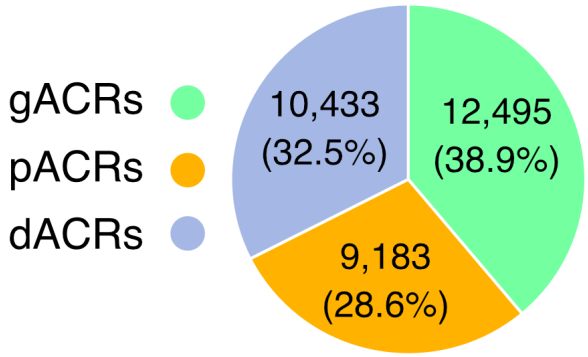

Figure 1. Genome-wide distribution of accessible chromatin regions (ACRs) in the maize genome. ATAC-seq data generated from maize leaf tissue was aligned to the maize AGPv4 reference genome and ACRs were categorized in relation to the maize AGPv4.38 annotated genes. gACRs are in or overlap with genes, pACRs fall within 2,000 bp of genes, dACRs are at least 2,000 bp from genes.

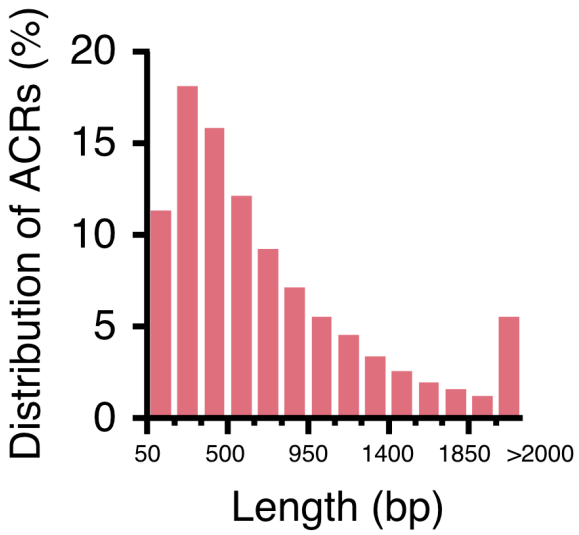

Figure 2. Length distribution of maize accessible chromatin regions (ACRs).

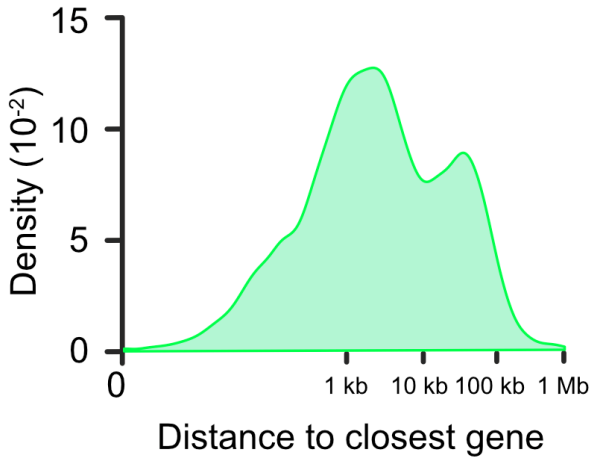

Figure 3. Kernel density of distance of accessible chromatin regions (ACRs) from the closest annotated genes. gACRs are excluded from this plot. The distances are $\log_2$ transformed.

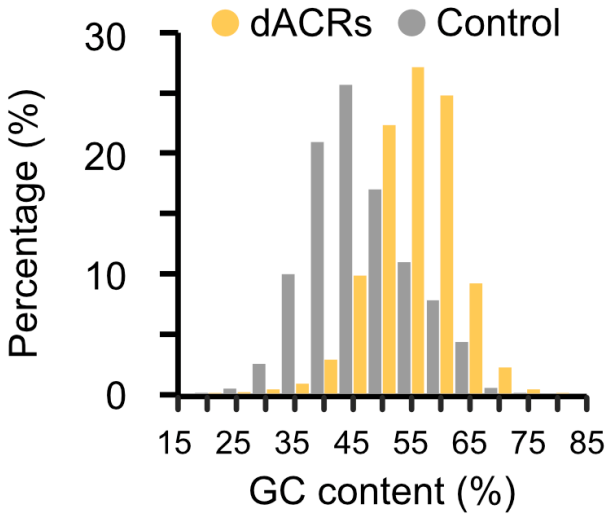

Figure 4. Distribution of GC content in dACRs and gene-distal uniquely mapping negative control regions.

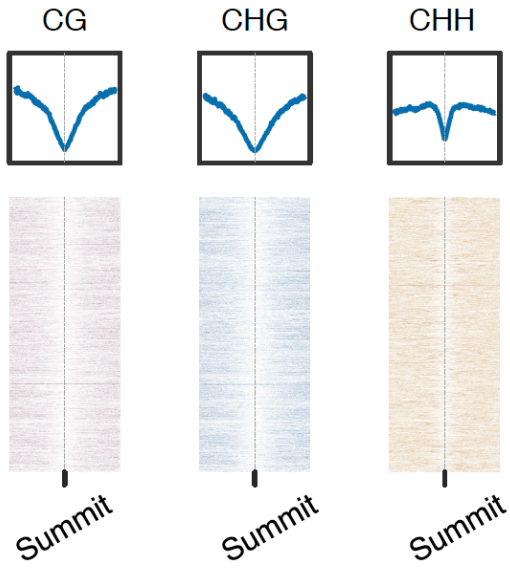

Figure 5. DNA methylation distribution of dACRs and flanking regions. DNA methylation (MethylC-seq) are aligned at the summits of dACRs and flanking 2kb regions. Weighted methylation levels are displayed.

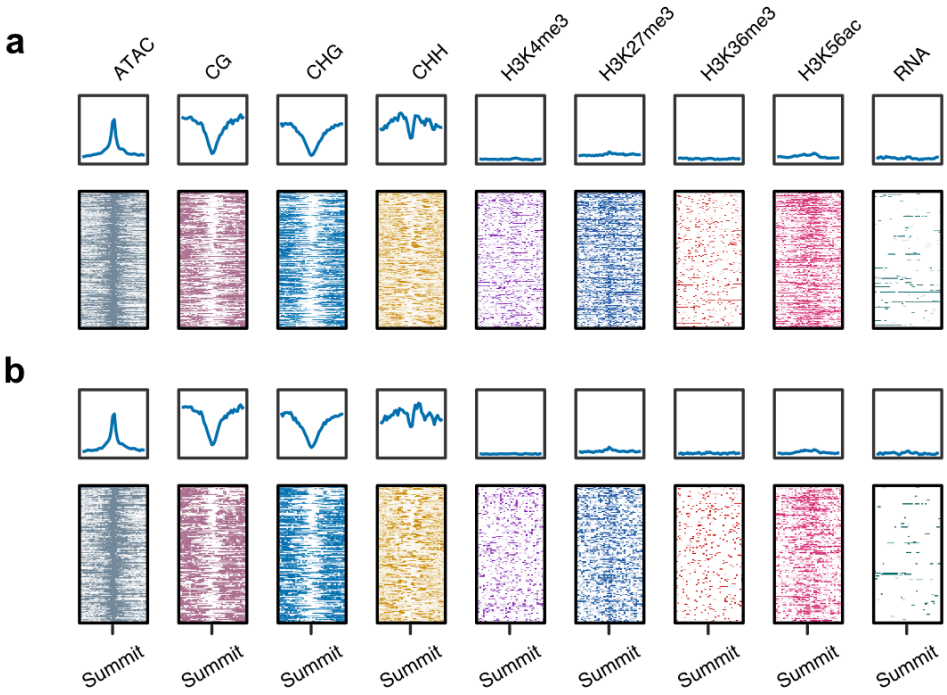

Figure 6. Chromatin attributes of dACRs in unmodified group. Chromatin accessibility (ATAC-seq), DNA methylation (MethylC-seq), Histone modifications (ChIP-seq), and transcriptional profile (RNA-seq) are aligned at the summits of dACRs and flanking 2kb regions. (**a**) Regions were identified by K-means clustering method. (**b**) Regions were identified by smoothing spline clustering method.

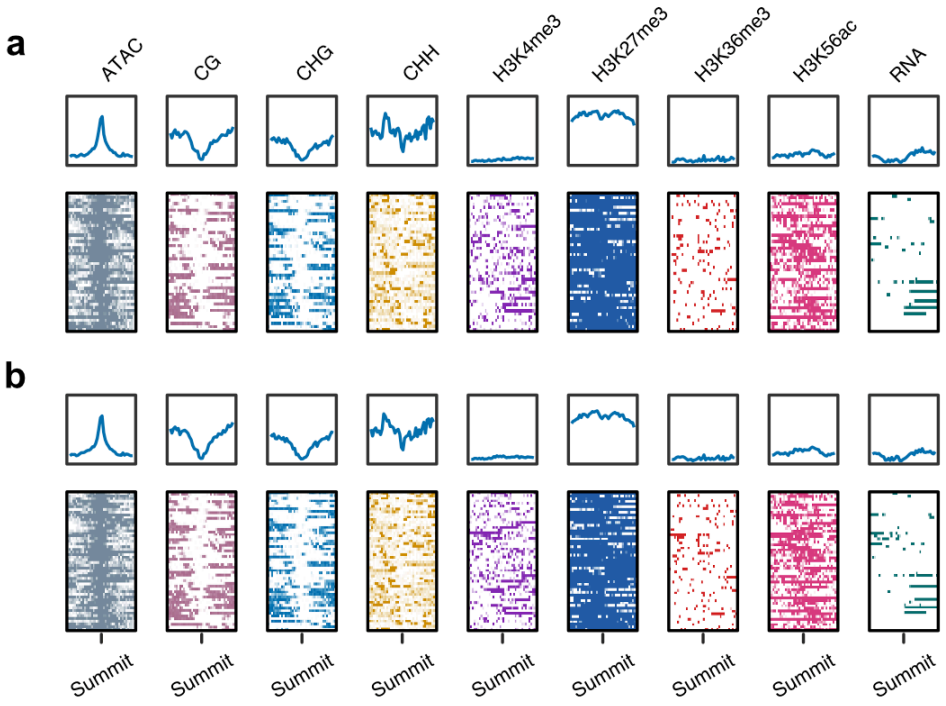

Figure 7. Chromatin attributes of dACRs in K27me3 group. Chromatin accessibility (ATAC-seq), DNA methylation (MethylC-seq), Histone modifications (ChIP-seq), and transcriptional profile (RNA-seq) are aligned at the summits of dACRs and flanking 2kb regions. (**a**) Regions were identified by K-means clustering method. (**b**) Regions were identified by smoothing spline clustering method.

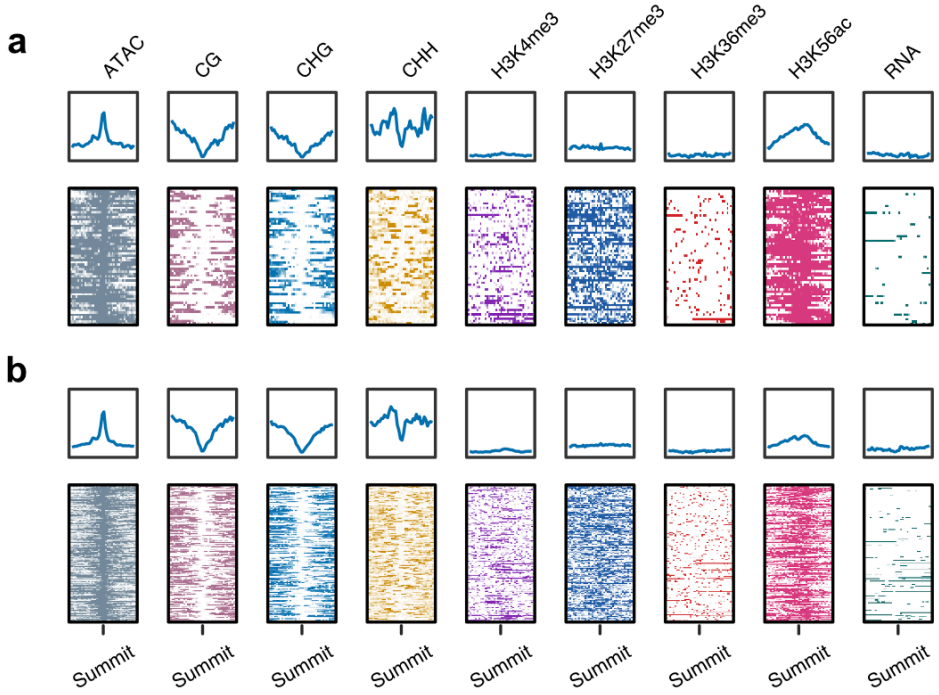

Figure 8. Chromatin attributes of dACRs in Kac group. Chromatin accessibility (ATAC-seq), DNA methylation (MethylC-seq), Histone modifications (ChIP-seq), and transcriptional profile (RNA-seq) are aligned at the summits of dACRs and flanking 2kb regions. (**a**) Regions were identified by K-means clustering method. (**b**) Regions were identified by smoothing spline clustering method.

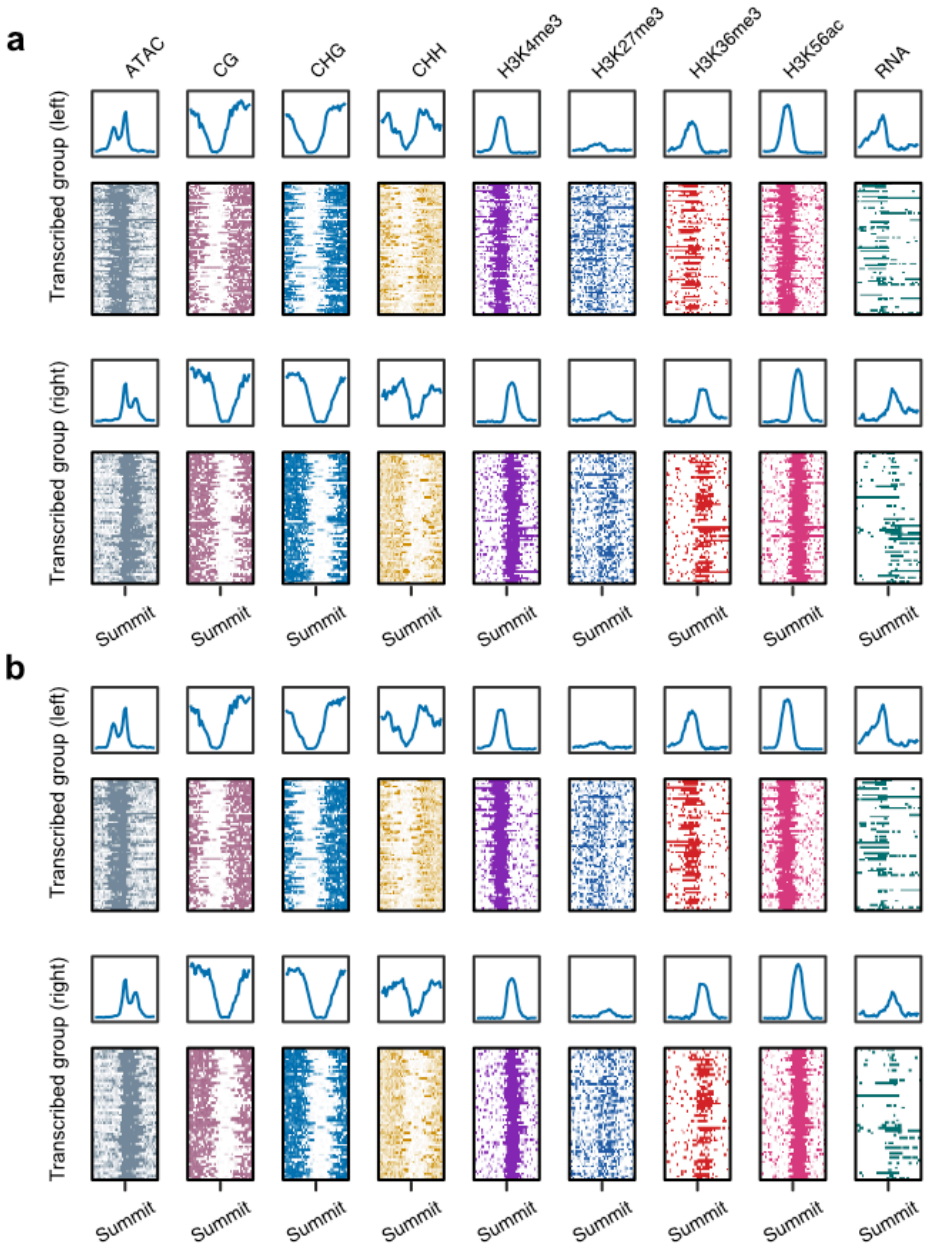

Figure 9. Chromatin attributes of dACRs in transcribed group. Chromatin accessibility (ATAC-seq), DNA methylation (MethylC-seq), Histone modifications (ChIP-seq), and transcriptional profile (RNA-seq) are aligned at the summits of dACRs and flanking 2kb regions. (**a**) Regions were identified by K-means clustering method. (**b**) Regions were identified by smoothing spline clustering method.

REFERENCES

Berger SL. 2002. Histone modifications in transcriptional regulation. *Curr. Opin. Genet. Dev.* **12**: 142-148.

Bolger AM et al. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114-2120.

Buenrostro JD et al. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**: 1213-1218.

Clark SJ et al. 1994. High sensitivity mapping of methylated cytosines. *Nucleic Acids Res.* **22**: 2990-2997.

Clark SJ et al. 2006. DNA methylation: bisulphite modification and analysis. *Nat. Protoc.* **1**: 2353-2364.

Cokus SJ et al. 2008. Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **452**: 215.

Craven P, Wahba G. 1978. Smoothing noisy data with spline functions. *Numerische mathematik* **31**: 377-403.

Du J et al. 2015. DNA methylation pathways and their crosstalk with histone methylation. *Nat. Rev. Mol. Cell Biol.* **16**: 519-532.

Frommer M et al. 1992. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. U.S.A.* **89**: 1827-1831.

Gu C, Ma P. 2005. Optimal smoothing in nonparametric mixed-effect models. *The Annals of Statistics* **33**: 1357-1379.

Hartigan JA, Wong MA. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C* **28**: 100-108.

Hayatsu H. 1976. Bisulfite modification of nucleic acids and their constituents. *Prog. Nucleic Acid Res. Mol. Biol.* **16**: 75-124.

Hayatsu H et al. 1970. Reaction of sodium bisulfite with uracil, cytosine, and their derivatives. *Biochemistry (Mosc).* **9**: 2858-2865.

Jain AK. 2010. Data clustering: 50 years beyond K-means. *Pattern Recog. Lett.* **31**: 651-666.

Ji L et al. 2014. Methylated DNA is over-represented in whole-genome bisulfite sequencing data. *Front Genet* **5**: 341.

Jiao Y et al. 2017. Improved maize reference genome with single-molecule technologies. *Nature* **546**: 524-527.

Johnson L et al. 2004. Mass spectrometry analysis of *Arabidopsis* histone H3 reveals distinct combinations of post-translational modifications. *Nucleic Acids Res.* **32**: 6511-6518.

Jones PA. 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**: 484-492.

Kaplan N et al. 2009. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458**: 362-366.

Kim D et al. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**: 357-360.

Langmead B et al. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**: R25.

Law JA, Jacobsen SE. 2010. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* **11**: 204-220.

Li H et al. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Lister R et al. 2008. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133**: 523-536.

Lu Z et al. 2017. Combining ATAC-seq with nuclei sorting for discovery of cis-regulatory regions in plant genomes. *Nucleic Acids Res.* **45**: e41.

Ma P et al. 2006. A data-driven clustering method for time course gene expression data. *Nucleic Acids Res.* **34**: 1261-1269.

Malone JH, Oliver B. 2011. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biol* **9**: 34.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17**: pp. 10-12.

Matzke MA, Mosher RA. 2014. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nat. Rev. Genet.* **15**: 394-408.

Niederhuth CE et al. 2016. Widespread natural variation of DNA methylation within angiosperms. *Genome Biol.* **17**: 194.

Park PJ. 2009. ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* **10**: 669-680.

Pertea M et al. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**: 290-295.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841-842.

Richmond TJ, Davey CA. 2003. The structure of DNA in the nucleosome core. *Nature* **423**: 145-150.

Schultz MD et al. 2015. Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* **523**: 212-216.

Schultz MD et al. 2012. 'Leveling' the playing field for analyses of single-base resolution DNA methylomes. *Trends Genet.* **28**: 583-585.

Schwarz G. 1978. Estimating the dimension of a model. *The annals of statistics* **6**: 461-464.

Shapiro R et al. 1973. Nucleic acid reactivity and conformation II. Reaction of cytosine and uracil with sodium bisulfite. *J. Biol. Chem.* **248**: 4060-4064.

Shapiro R et al. 1974. Deamination cytosine derivatives by bisulfite. Mechanism of the reaction. *J. Am. Chem. Soc.* **96**: 906-912.

Strahl BD, Allis CD. 2000. The language of covalent histone modifications. *Nature* **403**: 41-45.

Stroud H et al. 2013. Comprehensive analysis of silencing mutants reveals complex regulation of the *Arabidopsis* methylome. *Cell* **152**: 352-364.

Studer A et al. 2011. Identification of a functional transposon insertion in the maize domestication gene *tb1*. *Nat. Genet.* **43**: 1160-1163.

Tanurdzic M et al. 2008. Epigenomic consequences of immortalized plant cell suspension culture. *PLoS Biol.* **6**: 2880-2895.

Urich MA et al. 2015. MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. *Nat. Protoc.* **10**: 475-483.

Wagner EJ, Carpenter PB. 2012. Understanding the language of Lys36 methylation at histone H3. *Nat. Rev. Mol. Cell Biol.* **13**: 115-126.

Xu D, Tian Y. 2015. A comprehensive survey of clustering algorithms. *Annals of Data Science* **2**: 165-193.

Zhang X et al. 2009. Genome-wide analysis of mono-, di- and trimethylation of histone H3 lysine 4 in *Arabidopsis thaliana*. *Genome Biol.* **10**: R62.

Zhang X et al. 2007a. Whole-genome analysis of histone H3 lysine 27 trimethylation in Arabidopsis. *PLoS Biol.* **5**: e129.

Zhang X et al. 2007b. The Arabidopsis LHP1 protein colocalizes with histone H3 Lys27 trimethylation. *Nat. Struct. Mol. Biol.* **14**: 869-871.

Zhang Y et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**: R137.