Kansas State University Libraries

New Prairie Press

Conference on Applied Statistics in Agriculture 2010 - 22nd Annual Conference Proceedings

# MODELING DNA METHYLATION TILING ARRAY DATA

Gayla Olbricht
olbrichtg@mst.edu

Bruce A. Craig

R. W. Doerge

## Recommended Citation

Olbricht, Gayla; Craig, Bruce A.; and Doerge, R. W. (2010). "MODELING DNA METHYLATION TILING ARRAY DATA," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1061

# MODELING DNA METHYLATION TILING ARRAY DATA

Gayla R. Olbricht[1], Bruce A. Craig[1], and R. W. Doerge[1,2]

[1]Department of Statistics, Purdue University, West Lafayette, IN 47907, U.S.A.

[2]Department of Agronomy, Purdue University, West Lafayette, IN 47907, U.S.A.

**Abstract**: Epigenetics is the study of heritable changes in gene function that occur without a change in DNA sequence. It has quickly emerged as an essential area for understanding inheritance and variation that cannot be explained by the DNA sequence alone. Epigenetic modifications have the potential to regulate gene expression and may play a role in diseases such as cancer. DNA methylation is a type of epigenetic modification that occurs when a methyl chemical group attaches to a cytosine base on the DNA molecule. To better understand this epigenetic mechanism, DNA methylation profiles can be constructed by identifying all locations of DNA methylation in a genomic region (e.g. chromosome or whole-genome). Large-scale studies of DNA methylation are supported by microarray technology known as tiling arrays. These arrays provide high-density coverage of genomic regions through the unbiased, systematic selection of probes that are tiled across the regions. Statistical methods are employed to estimate each probe's DNA methylation status. Previous studies indicate that DNA methylation patterns of some organisms differ by genomic element (e.g., gene, transposon), suggesting that genomic annotation information may be useful in statistical analysis. In this work, a novel statistical model is proposed, which takes advantage of genomic annotation information that to date has not been effectively utilized in statistical analysis. Specifically, a hidden Markov model, which incorporates genomic annotation, is introduced and investigated through a simulation study and analysis of an *Arabidopsis thaliana* DNA methylation tiling array experiment.

## 1 Introduction

Understanding the factors that contribute to differences in observable traits (phenotypes) between individuals is a challenging task that has important implications in many areas of science and daily life. In agriculture, this is readily seen in efforts to produce crops or raise livestock with desirable characteristics, such as increased yield, resistance to disease and drought, or improved nutritional value. Historically, such traits have been investigated by studying the effects of the environment and genetics on the trait of interest. More recently, the field of epigenetics has flourished as an additional mechanism for explaining heritable phenotypic differences that cannot be explained by genetic information alone. Together, genetics and epigenetics can help in understanding heritable phenotypic variation. It is also important to be aware of the key differences in biological mechanisms that underlie these two modes of inheritance.

In the field of genetics, the Central Dogma of Molecular Biology (Crick, 1970) reveals how differences in the genetic material contained in DNA can lead to heritable variation in phenotypes between individuals with different DNA sequences. The Central Dogma states that DNA is transcribed to RNA, which is translated to protein (the fundamental unit of cellular function). Specifically, genes are subunits of DNA that encode a special class of RNA called messenger RNA (mRNA), which produces a chain of amino acids that form a protein (Griffiths et al., 2008). This process demonstrates how differences in DNA sequence can lead to the production of different proteins and thus introduce phenotypic variation.

The field of epigenetics focuses on understanding heritability that is not due to changes in the DNA sequence. Two common epigenetic modifications are DNA methylation and histone modifications, which involve the addition of chemical groups to the DNA or histone proteins without changing the DNA sequence itself (Figure 1). Epigenetic modifications can occur anywhere in the genome and have been shown to play a role in the regulation of gene expression (Zilberman et al., 2007) and development of cancer (Jones and Baylin, 2007). Epigenetics is currently an active area

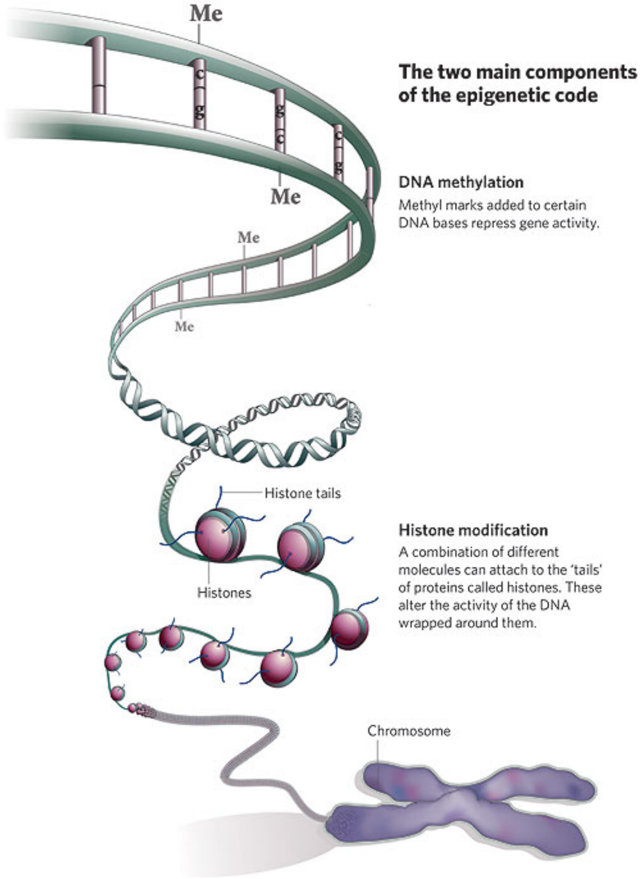
Me
c
g
g
c
c
g
Me
Me
Me
The two main components of the epigenetic code
DNA methylation
Methyl marks added to certain DNA bases repress gene activity.
Histone tails
Histones
Histone modification
A combination of different molecules can attach to the 'tails' of proteins called histones. These alter the activity of the DNA wrapped around them.
Chromosome

Figure 1: Illustration of the two main epigenetic modifications: DNA methylation and histone modifications. Image courtesy of Qiu (2006).

of research, as many epigenetic mechanisms are not well understood and a better understanding of these mechanisms can provide valuable insight into heritable differences that cannot be explained by changes in the DNA sequence.

In the 1990s, advances in technology made it possible to move from localized genetic and epigenetic studies to genome-wide investigations. The sequencing of DNA and identification of gene locations for entire genomes became a feasible task, with genome projects for over 1100 organisms being completed by September 2009 (GOLD: Genomes OnLine Database v 3.0, 2010). Online genomic annotation databases were subsequently created to store and make information about the location and function of genes and other genomic elements publicly available (Stein,

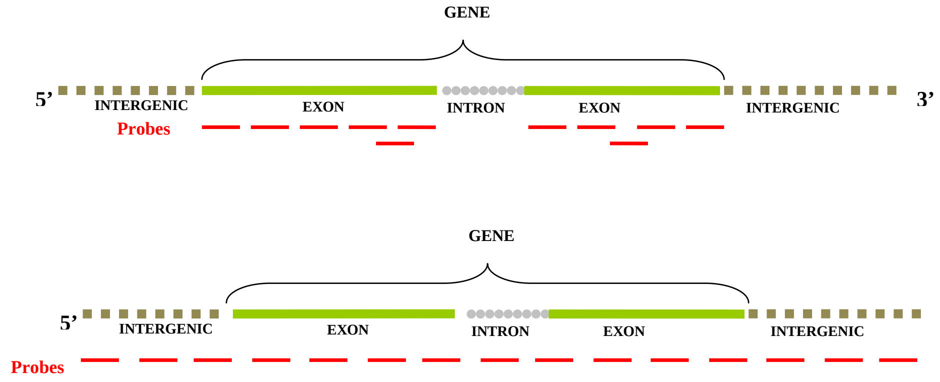

Figure 2: Top: Example of probes covering a genomic region on a gene expression microarray. Probes cover exons of genes and are designed to measure mRNA transcript abundance. Bottom: Example of probes covering a genomic region on a tiling array. Probes are placed from one end of the region to the other without regard to genomic annotation. As a result, exons and introns of genes, as well as intergenic regions are covered by probes. Images modified from Olbricht (2010).

2001). This wealth of information is currently available to aid genome-wide investigations of many types of biological mechanisms, including epigenetic modifications.

Along with genome projects, the development of microarray technology (Schena et al., 1995; Lockhart et al., 1996) provided the opportunity to investigate biological phenomena for a whole genome in a single experiment. Microarrays require the knowledge of DNA sequence information for the development of single-stranded probes, which are placed as targets on the array and have the potential to bind to a single-stranded mRNA or DNA sample via complementary base pair binding. Microarrays were initially used to study mRNA transcript abundance (i.e., gene expression level) by selecting probes from exons of genes (Figure 2, top) to determine which genes in a mRNA sample are active in making proteins (Schena et al., 1995). However, a unique type of microarray, called a tiling array, was soon developed to cover the whole genome (not just genes) through the systematic selection of probes from one end of a genomic region to the other (Figure 2, bottom). The dense, unbiased genomic coverage provided by tiling arrays make it possible to use microarray technology for epigenomic studies (Mockler and Ecker, 2005), in which statistical methods are employed to identify locations of epigenetic modifications across whole genomes.

This work focuses on the genome-wide study of DNA methylation using tiling array technology. Experimental methods and current statistical procedures applied in DNA methylation profiling studies are reviewed. Typically, results from these studies are visually connected to genomic annotation after the statistical analysis is complete to gain an understanding of the distribution of DNA methylation across the genome. In this work, we investigate the potential of more effectively utilizing the information available through genomic annotation databases by incorporating such data into statistical methods. Specifically, we propose that knowledge of which probes belong to which genomic regions can be valuable for statistical analysis and that integrating this genomic annotation information into statistical methods can help improve prediction of DNA methylation status. In particular, a hidden Markov model, which incorporates differences in DNA methylation patterns between gene and intergenic regions, is investigated through a simulation study and analysis of DNA methylation tiling array data generated from the model plant, *Arabidopsis thaliana*.

## 2 DNA Methylation Profiling with Tiling Arrays

DNA methylation is a type of epigenetic modification that typically occurs when a methyl group (Me) attaches to a cytosine (C) base on the DNA molecule (Figure 1). Although the addition of this chemical group does not alter the DNA sequence itself, it can have a profound impact on gene function. In mammals, DNA methylation typically occurs at sites where a cytosine is followed by a guanine (CG) in the $5' - 3'$ direction of the DNA sequence (Li and Bird, 2007). In plants, it can also occur at CNG and CNN (where N is one of the nucleotide bases adenine (A), cytosine (C), or thymine (T)) sites (Chan et al., 2005). DNA methylation has been shown to play an important role in many biological processes from embryonic development (Bird, 2002) to silencing of transposable elements (Slotkin and Martienssen, 2007), and has been linked to the development of human cancer (Jones and Baylin, 2007).

Since DNA methylation can vary between cell types and over time within an individual organism, determining the role of DNA methylation is often a complex task. A key to better understanding DNA methylation is to develop genome-wide profiles for different cell types by identifying all locations of DNA methylation in a genomic region. Tiling arrays enable such investigations due to their high-density, unbiased coverage that is essential for the study of DNA methylation, which can occur anywhere in the genome at specific cytosine sites. Tiling arrays have been successfully employed to evaluate DNA methylation status in many large-scale studies (Lippman et al., 2004; Zhang et al., 2006; Zilberman et al., 2007).

### 2.1 Experimental Methods

In DNA methylation profiling experiments, genomic DNA samples are collected and prepared with a treatment that allows DNA methylation to be measured with tiling arrays (Figure 3). As a first step in the sample preparation process, DNA collected from an individual is split into two subsamples and sheared into similar sizes. In one of these samples, a treatment such as bisulfite conversion, methylation sensitive restriction enzyme (e.g., McrBC) digestion, or methylcytosine immunoprecipitation, is applied to separate methylated from unmethylated DNA. No treatment is applied to the other sample, which serves as a control since it is representative of the total genomic DNA with both methylated and unmethylated DNA. Double-stranded DNA from both the treated and untreated samples are separated to single-stranded DNA and hybridized to tiling arrays (Beck and Rakyan, 2008). This process is repeated for additional biological replicates.

Statistical methods are needed to compare hybridization intensities between the treated and untreated samples for each probe to estimate whether the probe is methylated or not. For example, when the chemical treatment removes methylated DNA (e.g., McrBC digestion, Figure 3), then methylated probes are expected to have higher hybridization intensities in the untreated sample than in the treated sample, since the untreated sample retains methylated DNA and the treated sample does not. Note that probes are typically pre-processed via background correction, normalization, and log-transformation prior to implementation of a statistical model.

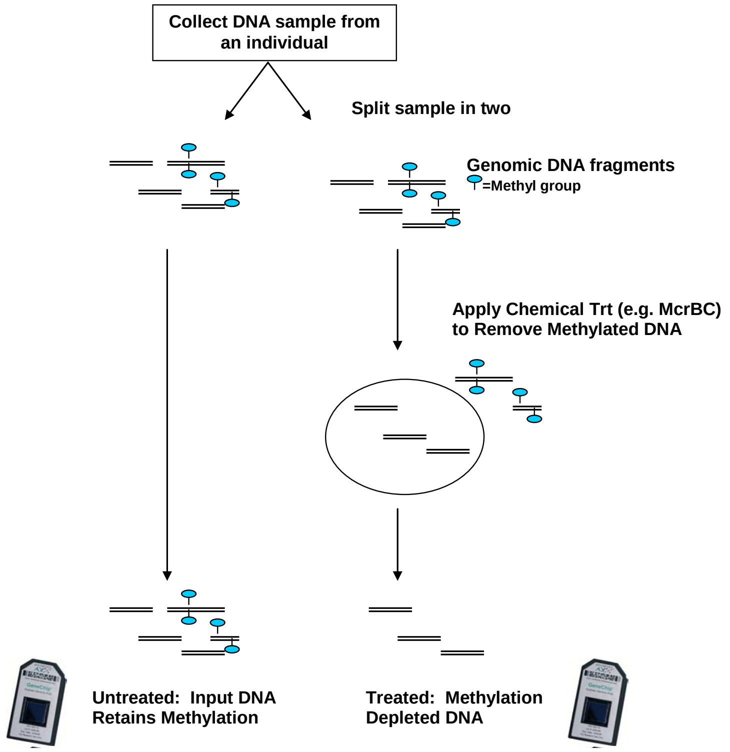
Collect DNA sample from an individual
Split sample in two
Genomic DNA fragments
=Methyl group
Apply Chemical Trt (e.g. McrBC) to Remove Methylated DNA
Untreated: Input DNA Retains Methylation
Treated: Methylation Depleted DNA

Figure 3: DNA sample preparation for DNA methylation profiling studies with tiling arrays. Here, DNA from one individual is split into two samples. Digestion with a methylation restriction enzyme (e.g. McrBC) is employed in one sample to remove methylated DNA. Single-stranded DNA from both samples is then hybridized to tiling arrays. Image courtesy of Olbricht (2010).

## 2.2 Current Statistical Methods

Data generated from DNA methylation profiling experiments present many statistical challenges due to the large number of probe-level tests (often millions), typically small number of biological replicates, dependency between neighboring probes, and experimental noise present in the data. One approach for determining whether a probe is significantly methylated is to employ an analysis of variance (ANOVA) model or conduct a paired $t$-test at each probe. Although several studies (Lippman et al., 2004; Martienssen et al., 2005; Vaughn et al., 2007) have successfully applied such models to identify DNA methylation status using tiling arrays, one issue that arises in this framework is the multiple testing problem. This issue is typically addressed by controlling the false discovery rate (FDR) at level $\alpha$ (Benjamini and Hochberg, 1995), which assumes the probe-specific hypothesis tests are independent. However, the linear ordering of probes across a genomic region makes this independence assumption questionable. Also, previous studies have shown that for many organisms, methylated probes tend to occur together in regions of dense methylation (Suzuki and Bird, 2008), further indicating that the DNA methylation status of a given probe may depend on neighboring probes.

Sliding window testing is an alternative statistical method that has been employed in tiling array studies to incorporate the potential dependency between neighboring probes (Cawley et al., 2004; Ji and Wong, 2005; Keles et al., 2006). Sliding window methods combine information from probes within a certain genomic distance of the probe being tested to calculate a test statistic for that probe. The test statistic and method of combining probes may differ for each proposed method. For example, Cawley et al. (2004) use all probes within a window of 1000 bases of the probe being tested to calculated a Wilcoxon rank sum statistic, while Keles et al. (2006) use a moving average of $t$-statistics. Although these methods take advantage of neighboring probe information, selecting an appropriate window size can be difficult and the multiple testing problem still remains an issue.

An alternative approach, which incorporates dependency among neighboring tiling array probes, is a hidden Markov model (HMM) (Rabiner, 1989). HMMs have been proposed and successfully applied in a variety of tiling array applications, including DNA methylation profiling experiments (Li et al., 2005; Ji and Wong, 2005; Du et al., 2006; Humburg et al., 2008; Yoo, 2008). In a hidden Markov model, a sequence of non-observable (hidden) random variables take on values in a set of finite states and form a first-order Markov chain. Although the states themselves are not directly observable, an observable output is available which is dependent on the hidden states. In the case of DNA methylation profiling experiments, the hidden states are the true methylation status of the probes (methylated or unmethylated) and the observed values are the intensity measurements obtained from the tiling array experiment (Figure 4).

HMM model parameters consist of initial probabilities ($\pi_i$), transition probabilities ($a_{ij}$), and the distribution parameters for the observations. The $\pi_i$ give the probability of the first probe being

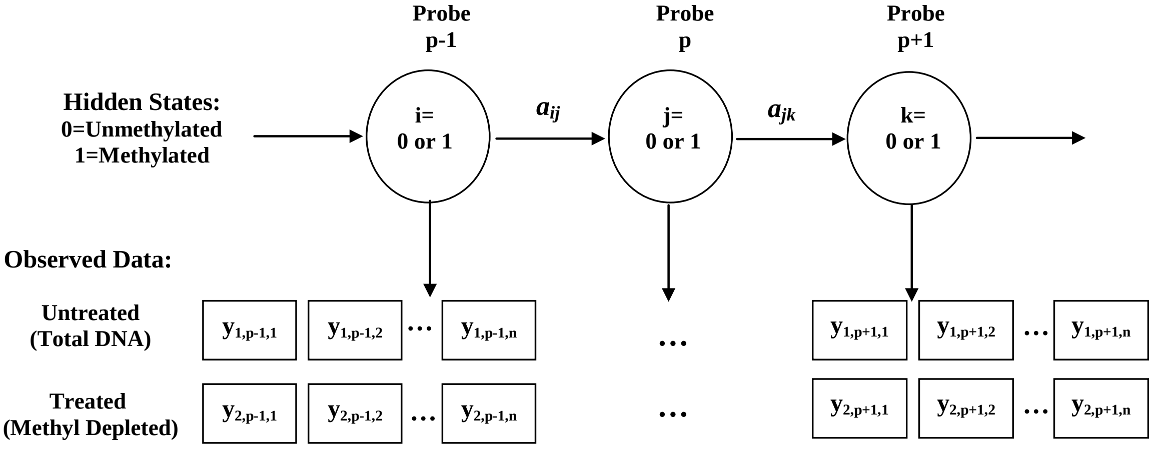

Figure 4: A hidden Markov model for DNA methylation profiling using tiling arrays. The circles represent the hidden states: 0 for unmethylated probes and 1 for methylated probes. The boxes represent the observations ($y_{lpr}$) where $l = \{1, 2\}$ is the sample type (untreated, treated), $p = \{1, ..., P\}$ is the probe, and $r = \{1, ..., n\}$ is the biological replicate. Arrows represent conditional dependencies. The hidden states for the probes follow a first-order Markov chain with transition probabilities $a_{ij}$ from probe $p - 1$ to probe $p$. The distribution of the observed data for each probe is conditionally dependent upon the hidden state at that probe. Image courtesy of Olbricht (2010).

in state $i$, while the $a_{ij}$ give the probability of moving from methylation status $i$ at probe $p - 1$ to methylation status $j$ at probe $p$. The observation probability distribution may differ according to the proposed method. Standard algorithms are available for HMMs, which can estimate the model parameters and the hidden states using information from all probes. Specifically, the Baum-Welch (BW) algorithm (Baum et al., 1970) calculates the maximum likelihood estimates for HMM model parameters, while the forward-backward (FB) algorithm (Baum et al., 1970; Baum, 1972) estimates the hidden states (i.e., DNA methylation status) for each probe.

## 2.3 Incorporating Genomic Annotation Information

The statistical methods described in Section 2.2 provide several options for determining the DNA methylation status of tiling array probes in a DNA methylation profiling experiment. Typically, online genomic annotation databases are employed to link probe position information to the location

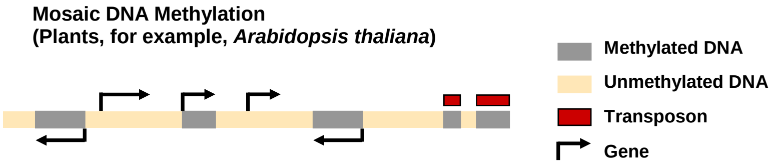

Figure 5: Example of mosaic DNA methylation, which occurs when densely methylated regions (grey) are interspersed with unmethylated or less densely methylated regions (yellow). In this example, genes (arrows) are either heavily methylated or completely unmethylated, and transposons (red) are methylated. Image courtesy of Olbricht (2010).

of different types of genomic elements (e.g., genes), thus determining which genomic element each probe represents. Results from the statistical analysis are then connected to genomic annotation to investigate patterns of DNA methylation according to different types of genomic elements. Using this strategy, it has been shown that different organisms show different overall DNA methylation patterns (Suzuki and Bird, 2008). Mammals often exhibit a global pattern, where DNA methylation is found at most CG sites throughout the genome. An exception is in groups of short regions called CpG islands, which are typically unmethylated in mammals. Some plants, such as maize, have high levels of DNA methylation, but others such as the model plant *Arabidopsis thaliana* display a mosaic DNA methylation pattern, where regions of dense methylation are interspersed with unmethylated regions (Figure 5) (Suzuki and Bird, 2008).

*Arabidopsis thaliana* was the first organism for which a genome-wide map of DNA methylation was constructed (Zhang et al., 2006; Zilberman et al., 2007), with almost 20% of the genome exhibiting dense DNA methylation. These studies suggest that dense DNA methylation typically occurs in transposons and inactive heterochromatin. In addition, over 30% of all genes are densely methylated in their transcribed regions, with transcription not generally suppressed by this gene body methylation (Zhang et al.,2006; Zilberman et al., 2007; Suzuki and Bird, 2008). This pattern of longer regions of DNA methylation in certain genomic regions (e.g., transposons, gene bodies) interspersed with unmethylated or less densely methylated regions (Figure 5), suggests that incorporating genomic annotation into statistical methods (rather than using this information solely

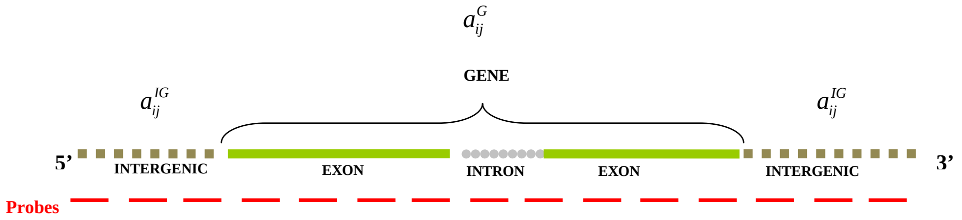

Figure 6: Incorporating genomic annotation into the HMM framework for DNA methylation tiling array experiments. Probes that correspond to gene regions can have different transition probabilities ($a_{ij}^{G}$) than probes in intergenic regions ($a_{ij}^{IG}$) to reflect different dependency patterns in those regions. Image courtesy of Olbricht (2010).

after the analysis is complete) may be valuable.

Here, we investigate a method for improving DNA methylation status prediction by using knowledge of genomic annotation in the statistical analysis of DNA methylation tiling array data. The hidden Markov model (HMM) framework offers a convenient way to model different DNA methylation patterns for different types of genomic elements through modifications to the transition probabilities. Current methods assume that transition probabilities are the same across the genomic region being investigated. To incorporate genomic annotation, probes in gene regions are allowed to have different transition probabilities ($a_{ij}^{G}$) than probes in intergenic regions ($a_{ij}^{IG}$) (Figure 6). Modifications are made to the forward-backward (FB) and Baum-Welch (BW) algorithms for parameter and hidden state estimation, which allow for differences in transition probabilities between genes and intergenic regions. The resulting model integrates the use of neighboring probe dependency with genomic annotation, while obtaining maximum likelihood estimates of HMM parameters. See (Olbricht, 2010) for further details on modifications to these algorithms.

## 3 Simulation Study

A simulation study is employed to investigate the importance of incorporating dependency between neighboring probes and utilizing genomic annotation in the HMM framework for DNA methylation profiling studies. Hidden states and observations are simulated for a genomic region of 2000 probes covering 20 genes. The initial state distribution is assumed to be $\pi = (0.5, 0.5)$ and transition probabilities are assumed to be different for gene regions ($a_{ij}^{G} = \left(\begin{smallmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \end{smallmatrix}\right)$) and intergenic regions ($a_{ij}^{IG} = \left(\begin{smallmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{smallmatrix}\right)$). Observed data are generated from a variety of different parameter settings of the following observation probability distribution:

$$\begin{aligned} \begin{pmatrix} y_{1pr} \\ y_{2pr} \end{pmatrix} &\sim N\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix} \right) \quad \text{Unmethylated Probes} \\ \begin{pmatrix} y_{1pr} \\ y_{2pr} \end{pmatrix} &\sim N\left( \begin{pmatrix} \mu_{11} \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix} \right) \quad \text{Methylated Probes.} \end{aligned} \tag{1}$$

These parameter settings are chosen so that in the unmethylated case, the untreated ($l = 1$) and treated ($l = 2$) means are equal and centered at zero, but in the methylated case there is a difference of ($\mu_{11}$) between the two means. Decreasing the magnitude of $\mu_{11}$ and increasing $\sigma$ should result in observed data in which state estimation is more difficult since the mean difference between untreated and treated samples will be smaller for the methylated case and the variation in the data larger. The value of $\rho$ is selected to allow for both a high and low level of correlation between samples from the same individual. All combinations of the following observation probability distribution parameter settings (2) are employed to simulate three biological replicates:

$$\mu_{11} = \{0.75, 1, 2\} \quad \sigma = \{1, 2\} \quad \rho = \{0.3, 0.7\}. \tag{2}$$

These data are simulated 1000 times. Averages over the three biological replicates are calculated for input into the forward-backward or the Baum-Welch algorithms.

The goals of this simulation study are two-fold. First, the results of a paired $t$-test conducted at each probe are compared to a HMM to evaluate the importance of modeling the probe dependency structure. Also, performance of a HMM which incorporates genomic annotation into hidden state estimation and a HMM that does not utilize this information are compared. Models are evaluated under the best case scenario when the model parameters are known, and thus no parameter estimation is required. For details on model evaluation when model parameters are unknown and the Baum-Welch algorithm is employed for parameter estimation, see Olbricht (2010). In this simulation, the following three models are compared.

**Independent Paired $t$-tests**: For each probe, a paired $t$-test is performed to determine whether the probe is methylated or not. Dependence between probes and genomic annotation are both ignored in this model.

**Unannotated HMM**: A hidden Markov model is employed with genomic annotation information ignored and transition probabilities assumed to be the same across the entire genomic region. The common transition probabilities are assumed to be the weighted average of $a_{ij}^{IG}$ and $a_{ij}^{G}$ (i.e., $a_{ij} = (\begin{smallmatrix} 0.87 & 0.13 \\ 0.13 & 0.87 \end{smallmatrix})$). The standard forward-backward (FB) algorithm is employed for hidden state estimation.

**Annotated HMM**: A hidden Markov model is employed with genomic annotation information incorporated into the HMM by assuming the gene has different transition probabilities ($a_{ij}^{G}$) than the intergenic region ($a_{ij}^{IG}$). The modified forward-backward (FB) algorithm which integrates the transition probability differences is employed for hidden state estimation.

Model performance is evaluated by calculating the proportion of estimated states that match the true states and averaging across the 1000 simulated datasets. Figure 7 shows the proportion of correctly predicted states for the three models across different settings (2) of the observation probability distribution (1). Across all settings, the HMMs shows a marked improvement over the independent paired $t$-tests and the annotated HMM outperforms the unannotated HMM. The difference in the magnitude of performance between the two HMMs increases as $\sigma$ increases and the mean difference between the untreated and treated samples ($\mu_{11}$) decreases, meaning that the

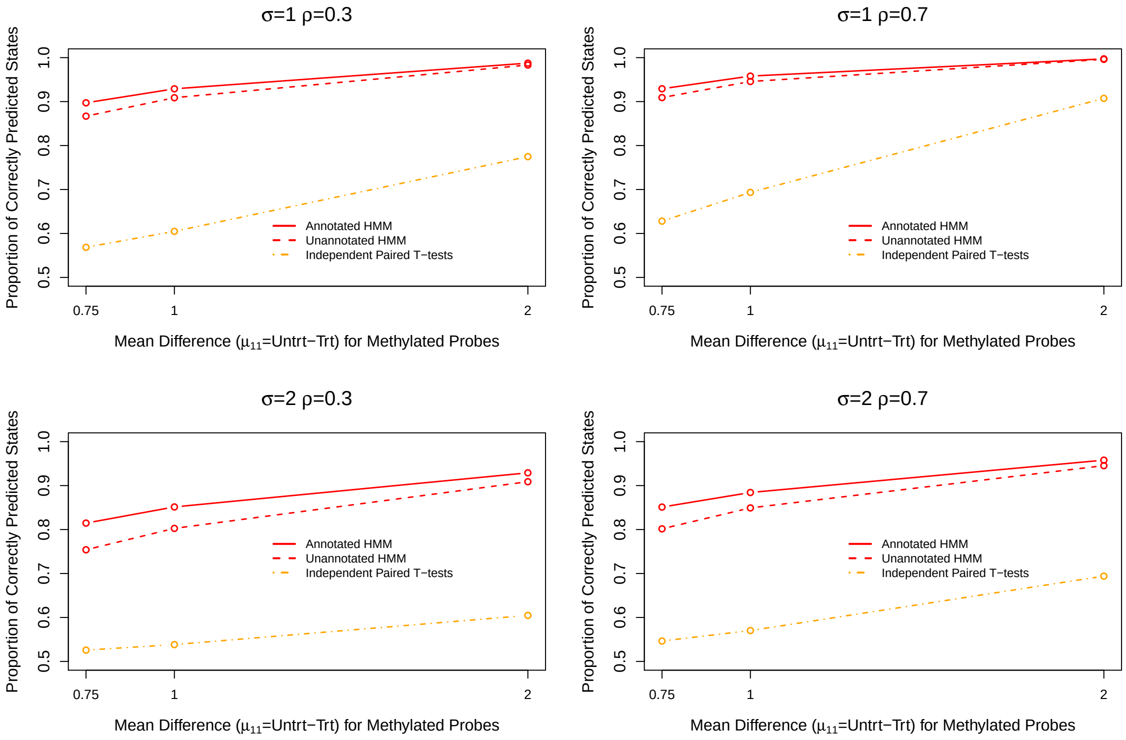

Figure 7: The proportion of states predicted correctly for the annotated and unannotated HMMs and the paired $t$-tests are plotted for each of the $\mu_{11}$ parameter settings of the observation probability distribution. Separate plots are shown for each combination of the $\sigma$ and $\rho$ parameters of the observation probability distribution. Image courtesy of Olbricht (2010).

annotated HMM can more accurately model noisy data and detect smaller mean differences than the unannotated HMM. The difference in model performance also appears to slightly increase as the correlation between the samples taken on the same subject ($\rho$) decreases, but this change in performance is only slight. Ultimately, these results show that it is important to incorporate the dependency structure between neighboring probes if it truly exists, as model performance of the independent paired $t$-tests was much worse than that of the HMMs. Further, incorporating genome annotation into HMM hidden state estimation improves prediction of DNA methylation status if there truly are differences in transition probabilities for genes and intergenic regions.

## 4 Data Analysis: *Arabidopsis thaliana* DNA Methylation Profiling Study of Chromosome 4

*Arabidopsis thaliana* is a small mustard plant that serves as the model organism for plants. In 2004, Lippman et al. custom-designed a tiling array to conduct a small scale study of epigenetic modifications in the heterochromatic knob on chromosome four (*hk4S*) of *Arabidopsis*. The heterochromatic knob is known to contain many transposons and repetitive DNA, which are often heavily methylated (Martienssen and Colot, 2001). Lippman et al. (2004) investigate DNA methylation, histone modifications, and gene expression in wild-type Columbia and a *ddm1* mutant of *Arabidopsis* all using the same tiling array platform. There are 1407 unique probes (each replicated two to four times) represented on the array that cover a 1.5 megabase (Mb) region centered on *hk4S*. Of the 1407 probes, 71.6% of them lie in gene regions, with an average of three probes per gene. The DNA methylation data obtained from wild-type Columbia *Arabidopsis* are further studied to gain a better understanding of the natural state of DNA methylation in this region. Unlike simulated data, the true underlying methylation status for each probe is unknown in these data. However, based on previous biological knowledge, it is expected that the heterochromatic knob will be more heavily methylated than the euchromatic regions surrounding it.

Lippman et al. (2004) employ the use of a methylation restriction enzyme (McrBC), as described in Section 2.1, to remove methylated DNA in the treated sample (Figure 3). Since two-color arrays are employed, both treated and untreated DNA samples are hybridized to the same array and a dye swap is performed. DNA samples are collected on two biological replicates, yielding a total of two arrays per individual and four arrays overall. To determine the DNA methylation status of each probe represented on the tiling array, Lippman et al. (2004) employ an ANOVA model with sample type (T) (treated or untreated), dye (D), array (A), probe (P) main effects and TP, DP, and AP interaction effects in the model. Yoo (2008) later reanalyze these data using the same ANOVA model, but updating the hypothesis tests to address issues specific to DNA methylation data. Specifically, Yoo (2008) conduct one-sided tests, utilizing a set of control probes that

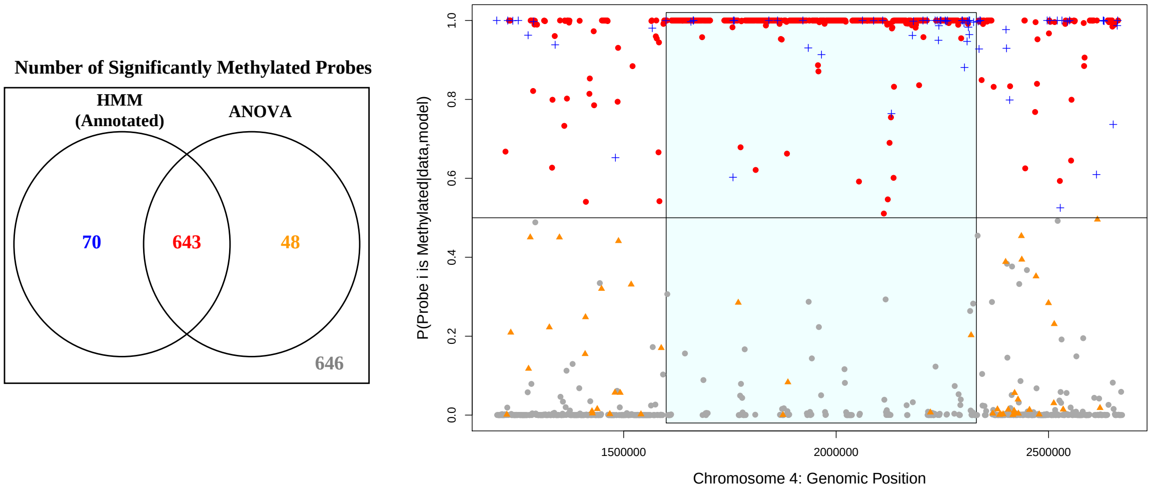

Figure 8: Left: Venn diagram comparing the number of significantly methylated probes identified by the annotated HMM and the ANOVA model. Right: The probability of each probe being in the methylated state (given the model parameters and data) plotted by the genomic location of each probe's start position. The colors of the symbols correspond to the colors in the Venn diagram, where red points (dots) are probes that are significantly methylated using both methods, blue points (crosses) are only found methylated in the annotated HMM, orange points (triangles) are only found methylated with ANOVA, and grey points (dots) are not identified as methylated in either method. The box highlights the heterochromatic knob region (1,600,000-2,330,000).

are known to be unmethylated as a reference. To address the multiple testing problem, the false discovery rate was controlled at $\alpha = 0.05$. Here, these data are further analyzed by applying the proposed hidden Markov model, which incorporates probe dependency and genomic annotation information. Results from the annotated HMM and the ANOVA employed by Yoo (2008) are compared.

Of the 1407 probes on the array, both the ANOVA model and the annotated HMM identify 643 as being significantly methylated (Figure 8, left). The ANOVA model identifies 48 significantly methylated probes that the annotated HMM does not, while the annotated HMM identifies 70 significantly methylated probes that the ANOVA model does not. Of probes in the heterochromatic knob, 74.9% are significantly methylated using ANOVA and 79.9% are identified as methylated using the annotated HMM. The high percentage of DNA methylation found by both methods in the knob region reaffirms the knowledge that heterochromatic DNA is heavily methylated and demonstrates the ability of both models to effectively detect this region of dense DNA methylation.

While the results between the two methods are similar, Figure 8 (right) highlights some of the differences. The probability of each probe being in the methylated state, given the model parameters and data, is calculated via the forward-backward algorithm in the HMM approach. This quantity is plotted by the genomic location of the start position of each probe. The annotated HMM identifies all points above 0.5 as being methylated. Note that the heterochromatic knob (highlighted in the box), contains many more methylated probes identified by both methods (red points) than the surrounding euchromatic region. There is also a lack of unmethylated probes (grey points) for both methods in that region. The significantly methylated probes identified by the ANOVA, but not the annotated HMM (orange points) are mostly located in the euchromatic region outside the knob. On the other hand, the annotated HMM identifies several methylated probes at the right end of the heterochromatic knob that the ANOVA model does not (blue points). Thus, although the true methylation status in these data is unknown, it is worth noting that the annotated HMM identifies more methylated probes than the ANOVA in the region where previous biological knowledge indicates methylation is occuring and less methylated probes outside that region.

Also, for the annotated HMM, further information can be gained by employing the modified Baum-Welch algorithm to obtain model parameter estimates. The parameter estimates for the transition probabilities for gene and intergenic regions are give below:

$$\hat{a}_{ij}^{IG} = \begin{pmatrix} 0.8649 & 0.1351 \\ 0.1348 & 0.8652 \end{pmatrix} \qquad \hat{a}_{ij}^{G} = \begin{pmatrix} 0.8620 & 0.1380 \\ 0.1337 & 0.8663 \end{pmatrix}$$

These parameter estimates indicate that the transition probabilities for genes and intergenic regions are similar for this region of the genome (i.e, the probability of staying in the same state is $\sim 0.86$). Although the differences in transition probabilities are small in magnitude, employing the annotated HMM allows the direct investigation of such patterns through statistical analysis that is not possible by previous methods.

## 5 Summary

DNA methylation is a type of epigenetic modification that plays an important role in many different biological processes and is one mechanism for establishing heritable phenotypic differences that cannot be explained by a change in the DNA sequence. DNA methylation profiling studies can offer valuable insight into this epigenetic mechanism by identifying the distribution of DNA methylation across a genomic region (e.g., chromosome or whole genome). DNA methylation profiling is accomplished by using a tiling microarray, which offers unbiased coverage of entire genomic regions through the sequential selection of probes from one end of the region to the other. Statistical methods, such as ANOVA, sliding window tests, or hidden Markov models, are employed to determine the DNA methylation status of each probe.

While previous statistical methods have been successfully employed to identify locations of DNA methylation in tiling array experiments, none of the current methods take advantage of the genomic annotation information that are available in online databases in the statistical analysis. In this work, a hidden Markov model (HMM) which incorporates genomic annotation information by modeling differences in transition probabilities between genes and intergenic regions is introduced. The annotated HMM is successfully applied to both simulated and real data, with results indicating that incorporation of genomic annotation information into a HMM framework is beneficial in predicting DNA methylation status.

While this work focuses on the breakdown of genomic annotation into genes and intergenic regions, it may be worthwhile to consider other types of genomic annotation (e.g., locations of transposons) for incorporation into a HMM for DNA methylation profiling studies. The methods proposed here can be extended to include more than two sets of transition probabilities for multiple types of genomic elements. Also, although the methods here are designed for tiling arrays, a newer type of technology, referred to as next-generation sequencing (NGS), has become a popular way to study many types of biological phenomena, including DNA methylation. Investigating how our methods can be extended to NGS studies will be an important endeavor in advancing this work.

## 6 Acknowledgements

We would like to thank Mary Ellen Bock, Michelle Lacey, Alan Qi, and the RWD research group for the helpful discussions and valuable feedback provided during the course of this work. We specifically thank Tilman Achberger for offering helpful comments on this manuscript. We are also grateful to Doug Crabill and My Troung for their computing expertise, as well as the Department of Statistics at Kansas State University for assistance with travel funding to GRO. This work is supported by a NSF Plant Genome grant (DBI-0733857) in part to RWD.

## References

Baum, L., T. Petrie, G. Soules, and N. Weiss (1970). A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics 41*, 164–171.

Baum, L. E. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities 3*, 1–8.

Beck, S. and V. K. Rakyan (2008). The methylome: approaches for global DNA methylation profiling. *Trends in Genetics 24*, 231–237.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological) 57*, 289–300.

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

Cawley, S., S. Bekiranov, H. H. Ng, P. Kapranov, E. A. Sekinger, D. Kampa, A. Piccolboni, V. Sementchenko, J. Cheng, A. J. Williams, R. Wheeler, B. Wong, J. Drenkow, M. Yamanaka, S. Patel, S. Brubaker, H. Tammana, G. Helt, K. Struhl, and T. R. Gingeras (2004). Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell 116*, 499–509.

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

Crick, F. (1970). Central dogma of molecular biology. *Nature 227*, 561–563.

Du, J., J. Rozowsky, J. O. Korbel, Z. D. Zhang, T. E. Royce, M. E. Schultz, M. Snyder, and M. Gerstein (2006, December). A supervised hidden Markov model framework for efficiently segmenting tiling array data in transcriptional and ChIP-chip experiments: systematically incorporating validated biological knowledge. *Bioinformatics 22*(24), 3016–3024.

GOLD: Genomes OnLine Database v 3.0 (2010). http://www.genomesonline.org/.

Griffiths, A. J., S. R. Wessler, R. C. Lewontin, and S. B. Carroll (2008). *Introduction to Genetic Analysis*. W.H. Freeman and Company.

Humburg, P., D. Bulger, and G. Stone (2008, August). Parameter estimation for robust HMM analysis of ChIP-chip data. *BMC Bioinformatics 9*, 343.

Ji, H. and W. H. Wong (2005, September). Tilemap: create chromosomal map of tiling array hybridizations. *Bioinformatics 21*(18), 3629–3636.

Jones, P. A. and S. B. Baylin (2007). The epigenomics of cancer. *Cell 128*, 683–692.

Keles, S., M. J. V. D. Laan, S. Dudoit, and S. E. Cawley (2006). Multiple testing methods for ChIP-chip high density oligonucleotide array data. *Journal of Computational Biology 13*, 579–613.

Li, E. and A. Bird (2007). DNA methylation in mammals. In *Epigenetics*. Cold Spring Harbor Laboratory Press.

Li, W., C. A. Meyer, and X. S. Lu (2005, June). A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics 21*(Suppl 1), i274–i282.

Lippman, Z., A.-V. Gendrel, M. Black, M. W. Vaughn, N. Dedhia, W. R. McCombie, K. Lavine, V. Mittal, B. May, K. D. Kasschau, J. C. Carrington, R. W. Doerge, V. Colot, and R. Martienssen (2004). Role of transposable elements in heterochromatin and epigenetic control. *Nature 430*, 471–476.

Lockhart, D. J., H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittman, C. W. Wang, M. Kobayashi, H. Horton, and E. L. Brown (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology 14*, 1675–1680.

Martienssen, R. A. and V. Colot (2001). DNA methylation and epigenetic inheritance in plants and filamentous fungi. *Science 293*, 1070–1074.

Martienssen, R. A., R. Doerge, and V. Colot (2005). Epigenomic mapping in Arabidopsis using tiling microarrays. *Chromosome Research 13*, 299–308.

Mockler, T. C. and J. R. Ecker (2005). Applications of DNA tiling arrays for whole-genome analysis. *Genomics 85*, 1–15.

Olbricht, G. R. (2010). *Incorporating genome annotation in the statistical analysis of genomic and epigenomic tiling array data*. Ph.D. dissertation, Purdue University, West Lafayette, IN USA.

Qiu, J. (2006). Epigenetics: unfinished symphony. *Nature 441*, 143–145.

Rabiner, L. R. (1989, February). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE 77*(2), 257–286.

Schena, M., D. Shalon, R. W. Davis, and P. O. Brown (1995). Quantitative monitering of gene expression patterns with a complementary DNA microarray. *Science 270*, 467–470.

Slotkin, R. K. and R. Martienssen (2007). Transposable elements and the epigenetic regulation of the genome. *Nature Reviews Genetics 8*, 272–285.

Stein, L. (2001). Genome annotation: from sequence to biology. *Nature Reviews Genetics 2*, 493–503.

Suzuki, M. M. and A. Bird (2008). DNA methylation landscapes: provocative insights from epigenetics. *Nature Reviews Genetics 9*, 465–476.

Vaughn, M. W., M. Tanurdzic, Z. Lippman, H. Jiang, R. Carrasquillo, P. D. Rabinowicz, N. Dedhia, W. R. McCombie, N. Agier, A. Bulski, V. Colot, R. Doerge, and R. A. Martienssen (2007). Epigenetic natural variation in Arabidopsis thaliana. *PLoS Biology 5*, e174.

Yoo, S.-Y. (2008). *Statistical methods for integrating epigenomic results*. Ph.D. dissertation, Purdue University, West Lafayette, IN USA.

Zhang, X., J. Yazaki, A. Sundaresan, S. Cokus, S. W.-L. Chan, H. Chen, I. R. Henderson, P. Shinn, M. Pellegrini, S. E. Jacobsen, and J. R. Ecker (2006). Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis. *Cell 126*, 1189–1201.

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*, 61–69.