01 Jan 2010

Incorporating Genome Annotation in the Statistical Analysis of Genomic and Epigenomic Tiling Array Data

Gayla R. Olbricht
Missouri University of Science and Technology, olbrichtg@mst.edu

Graduate School ETD Form 9
(Revised 12/07)

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Gayla Olbricht

Entitled Incorporating Genome Annotation in the Statistical Analysis of Genomic and Epigenomic Tiling Array Data

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Bruce A. Craig, Chair	Rebecca W. Doerge
Mary Ellen Bock	Michelle R. Lacey
Yuan (Alan) Qi	

To the best of my knowledge and as understood by the student in the *Research Integrity and Copyright Disclaimer (Graduate School Form 20)*, this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Approved by Major Professor(s): Rebecca W. Doerge

Bruce A. Craig

Approved by: Mary Ellen Bock (Head of the Graduate Program) 06/01/2010 (Date)

Graduate School Form 20
(Revised 1/10)

PURDUE UNIVERSITY
GRADUATE SCHOOL

Research Integrity and Copyright Disclaimer

Title of Thesis/Dissertation:

Incorporating Genome Annotation in the Statistical Analysis of Genomic and Epigenomic Tiling Array Data

For the degree of Doctor of Philosophy

I certify that in the preparation of this thesis, I have observed the provisions of *Purdue University Teaching, Research, and Outreach Policy on Research Misconduct (VIII.3.1)*, October 1, 2008.*

Further, I certify that this work is free of plagiarism and all materials appearing in this thesis/dissertation have been properly quoted and attributed.

I certify that all copyrighted material incorporated into this thesis/dissertation is in compliance with the United States' copyright law and that I have received written permission from the copyright owners for my use of their work, which is beyond the scope of the law. I agree to indemnify and save harmless Purdue University from any and all claims that may be asserted or that may arise from any copyright violation.

Gayla Olbricht

Printed Name and Signature of Candidate

07/14/2010

Date (month/day/year)

*Located at http://www.purdue.edu/policies/pages/teach_res_outreach/viii_3_1.html

INCORPORATING GENOME ANNOTATION IN THE STATISTICAL ANALYSIS OF GENOMIC AND EPIGENOMIC TILING ARRAY DATA

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Gayla R. Olbricht

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

August 2010

Purdue University

West Lafayette, Indiana

To my parents, who have provided a constant source of love and encouragement throughout my life, and to my husband Josh, whose loving support and belief in my potential to succeed continues to motivate me every day.

ACKNOWLEDGMENTS

During my time as a graduate student in the Department of Statistics at Purdue University, I have had the privilege of being surrounded by a wealth of talented, dedicated, and supportive individuals that have made my experience here exceptional. First and foremost, I would like to thank my two co-advisors, Professors Rebecca W. Doerge and Bruce A. Craig for their support, encouragement, direction, and friendship during my time as a Ph.D. student. I am very grateful to Rebecca for believing in me and offering a supportive ear from my first year of graduate school. The environment she creates for the students in her research group has helped me develop a breadth of professional skills and has exposed me to many new areas of scientific research. Her dedication to her students and profession is inspiring. I would like to thank Bruce for his helpful guidance, starting with my time as a consultant in the Statistical Consulting Service. He has always challenges me to think about things in new ways and I am grateful for his invaluable support.

I would also like to gratefully acknowledge my committee members, Professors Alan Qi, Mary Ellen Bock, and Michelle Lacey for their insights and helpful discussions. A special thanks to former department head Mary Ellen Bock, who has fostered an atmosphere within the Department of Statistics conducive to developing sound statistical skills, while encouraging respectful and positive interactions between faculty, students, and staff. I would like to thank the staff members in the department for their kindness and helpfulness as I navigated through graduate school. I would especially like thank Regina Becker, Teena Erwin, and Cheryl Crabill for generously giving of their time when I was a leader in the Statistics in the Community (STATCOM) program. I am also very grateful to system administrators, Doug Crabill and My Truong, who have helped me on numerous occasions with their invaluable computing expertise.

I am grateful to Dr. Jolena Waddell and Professor Christopher Bidwell in the Animal Sciences Department who provided my first opportunity as a consultant to learn about microarray data analysis. Thanks to Jolena for her friendship and teaching me many biological concepts about gene expression data. I would like to acknowledge Professor Stanton Gelvin and Dr. Nagesh Sardesai in the Biological Sciences Department for collaborating with me on a tiling array analysis and allowing me to use their data in Chapter 3 of this dissertation. I am also grateful to Dr. Jody Riskowski and Jennifer Wilson, whose creative insights during my time as a GK-12 Fellow at Wea Ridge Middle School have helped me become a better teacher. I am also indebted to Professor George Mathew at Missouri State University for encouraging me to pursue a career in statistics and for providing helpful advice along the way.

The friendships I have developed as a graduate student have made my time here rewarding and enjoyable. In particular, I am fortunate to have had the opportunity to work with current and past members of the RWD research group: Tilman Achberger, Lingling An, Paul Livermore Auer, Doug Baumann, Martina Muehlbach Bremer, Riyan Cheng, Kyunga Kim, Alex Lipka, Cherie Ochsenfeld, Andrea Rau (and the UWN), Sanvesh Srivastava, and Suk-Young Yoo. The camaraderie and fruitful discussions with this group has been one of the highlights of my time as a graduate student. I am also grateful for the friendship of Tina Alexander, Nilupa Gunaratna, Amy Wozniak, Andrew Lewandowski, and Amanda Bean who I have had the privilege of sharing valuable study time, STATCOM experiences, and many fond memories.

Finally, I would like to express my heartfelt gratitude for the love and support of my family, which has always been a source of strength and inspiration for me. I would especially like to thank my parents, Randal and Linda Hobbs, and my parents-in-law, Ted and Barbara Olbricht, for their love, patience, and encouragement, which have been particularly uplifting during my time in graduate school. Thanks to my sister, Rhonda, for her friendship throughout life. To my husband Josh, thank you for taking this adventure in life with me. Your companionship and unwaivering belief in me has inspired me more than you will ever know.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

ABSTRACT

Olbricht, Gayla R. Ph.D., Purdue University, August 2010. Incorporating Genome Annotation in the Statistical Analysis of Genomic and Epigenomic Tiling Array Data. Major Professors: R.W. Doerge and Bruce A. Craig.

A wealth of information and technologies are currently available for the genome-wide investigation of many types of biological phenomena. Genomic annotation databases provide information about the DNA sequence of a particular organism and give locations of different types of genomic elements, such as the exons and introns of genes. Microarrays are a powerful type of technology that make use of DNA sequence information to investigate different types of biological phenomena on a genome-wide level. Tiling arrays are a unique type of microarray that provide unbiased, high-density coverage of a genomic region, making them well suited for many applications, such as the mapping of transcription and the profiling of epigenetic mechanisms that can occur anywhere in the genome. Epigenetic mechanisms, such as DNA methylation and histone modifications, are important for understanding heritable changes in genome function that cannot be explained by a change in the DNA sequence alone.

In this work, statistical approaches for both gene expression and DNA methylation tiling array data are investigated. The proposed methods take advantage of the genomic annotation that are available and that to date have not been effectively utilized in current statistical methods. For gene expression data, an initial bioinformatic step, prior to differential expression analysis, is proposed for the purpose of filtering out probes that are biologically irrelevant. For DNA methylation data, a hidden Markov model, which allows for different transition probabilities between gene and intergenic regions is developed in an effort to improve the predicted locations of DNA methylation across the genome. These methods are investigated through simulation studies and real data analyses.

1. INTRODUCTION

1.1 Basics of Genetics

Understanding the factors that influence observable characteristics (phenotypes) of an organism is an important and complex task that has profound implications in agriculture, medicine, and many other areas of science. For example, determining why some plants of a species endure during periods of drought while others die, or why one identical twin develops a disease and the other does not, are just two of the many questions that could be addressed in understanding phenotypes. Three key components that can affect phenotype are genetic information, epigenetic modifications, and environmental conditions. The first of these components is studied in the field of genetics, which explores how hereditary information contained in deoxyribonucleic acid (DNA) is organized, passed from parent to offspring, and expressed phenotypically.

1.1.1 DNA Structure and Central Dogma of Molecular Biology

The science of genetics began with Gregor Mendel's work on inheritance of certain traits in the pea plant in the mid-1800s. The more recent discovery by Watson and Crick (1953) of the structure of the genetic material, deoxyribonucleic acid (DNA), has allowed a more detailed investigation of inheritance at the molecular level. DNA is a biological molecule present in the cells of an organism, which contains the genetic material needed to pass information from generation to generation. The DNA molecule has a double helix structure (Watson and Crick, 1953), which consists of two complementary strands composed of subunits called nucleotides. Each nucleotide contains a deoxyribose sugar and a phosphate group, which form the backbone, along

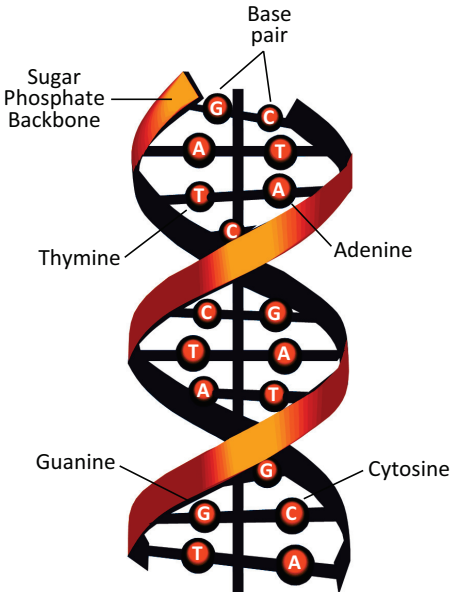

Figure 1.1. Structure of the DNA molecule. Image courtesy of Rau (2010).

with one of four nitrogenous bases: adenine (A), thymine (T), guanine (G), and cytosine (C). Each base pairs via a hydrogen bond with a base on the opposite strand in a complementary way, so that adenine always pairs with thymine and guanine always pairs with cytosine (Figure 1.1).

In eukaryotic organisms, DNA is found in the nucleus of each cell and is subdivided into units called chromosomes. Each chromosome consists of subunits of DNA called genes. Genes contain hereditary information needed to encode proteins, the fundamental unit of cellular function. The transfer of information from DNA to proteins is described in the Central Dogma of molecular biology (Crick, 1970), which states that DNA is transcribed to ribonucleic acid (RNA) which is translated to protein. Transcription occurs when one of the DNA strands serves as a template for creating a special class of RNA called messenger RNA (mRNA). Messenger RNA is a single-stranded complementary copy of the DNA strand with thymine (T) replaced by the nitrogenous base uracil (U) and a ribose sugar rather than deoxyribose in its backbone. During translation, mRNA produces a chain of amino acids that form a protein.

Genes are further broken down into subunits called exons (expression regions) and introns (intervening regions), with intergenic regions between genes. Exons are segments of the gene that encode parts of proteins, while introns are pieces that separate exons but do not encode parts of proteins. During RNA splicing, introns are removed and exons are joined together to create a mature mRNA transcript that can be translated to a protein. Variation in this splicing process can occur via alternative splicing in which different combinations of exons are spliced together to create multiple forms of the mRNA transcript. Alternative splicing makes it possible for different proteins to be produced by the same gene (Griffiths et al., 2008).

For organisms to function properly, genes must be regulated so that at a given time and specific cell type, only a subset of the genes are actively encoding proteins needed for a certain cellular function. Gene regulation can occur at many levels from transcription to post-translation. However, most regulation is thought to take place at the transcriptional level. A gene is referred to as expressed when it is active in making protein. Since a gene must first be transcribed to mRNA before being translated to a protein, the mRNA transcript contains the information needed to determine which genes are being expressed in a cell. By measuring the abundance of mRNA transcripts present for a particular gene, the expression level of that gene can be quantified (Griffiths et al., 2008).

1.1.2 Genetic Variation

It is evident through the Central Dogma that differences in the nucleotide sequence of genes can result in the production of different proteins, leading to phenotypic variation between individuals with different DNA sequences. The process of recombination during meiosis is one such source of genetic variation. During meiosis, sex cells are formed whose genetic material will be passed on to offspring upon fertilization. As part of this process, genetic material is often exchanged during recombination between homologous chromosomes, which contain the same genes but not necessarily the same

version of each gene. These recombination events can lead to a unique combination of nucleotides in the sex cells to be passed on to offspring (Griffiths et al., 2008).

Mutations are another source of genetic variation that occur when the DNA sequence is altered. Typically, mutations within a gene involve the alteration of one, or a few, nucleotide base pairs. This can occur when one nucleotide base is replaced by a different nucleotide base (substitution) or when one or more nucleotide base pairs are added to (insertion) or removed from (deletion) the DNA sequence. Mutations may occur for a variety of reasons, such as errors in the DNA replication process or environmental exposures such as certain chemical agents or radiation. However, the cell has a sophisticated set of DNA repair mechanisms in place that corrects most of these mutations. When this mechanism fails, the severity of the effect of mutations on the phenotype varies depending on the type of mutation, where it occurs within the gene, and how it effects the protein products of the mutated gene. In severe cases, mutations in sex cells can lead to inherited genetic disorders and may cause cancer when present in somatic cells (Griffiths et al., 2008).

1.2 Epigenetics

While the field of genetics reveals how differences in DNA sequence can lead to heritable variation in phenotype, the field of epigenetics seeks to understand heritability that is not due to changes in the DNA sequence. The term "epigenetics" was first introduced by Waddington (1942), who rooted the definition in concepts from epigenesis to be the study of how genotypes bring about phenotypes in the developmental process. Since that time, both the definition and field have evolved (Holliday, 2006; Bird, 2007; Berger et al., 2009), with the term "epigenetics" now commonly referring to the study of heritable changes in gene function that cannot be explained by changes in DNA sequence (Russo et al., 1996). The two main epigenetic mechanisms are DNA methylation and histone modifications, which involve the addition of chemical marks to the DNA or histone proteins (Figure 1.2). (Jones and Baylin,

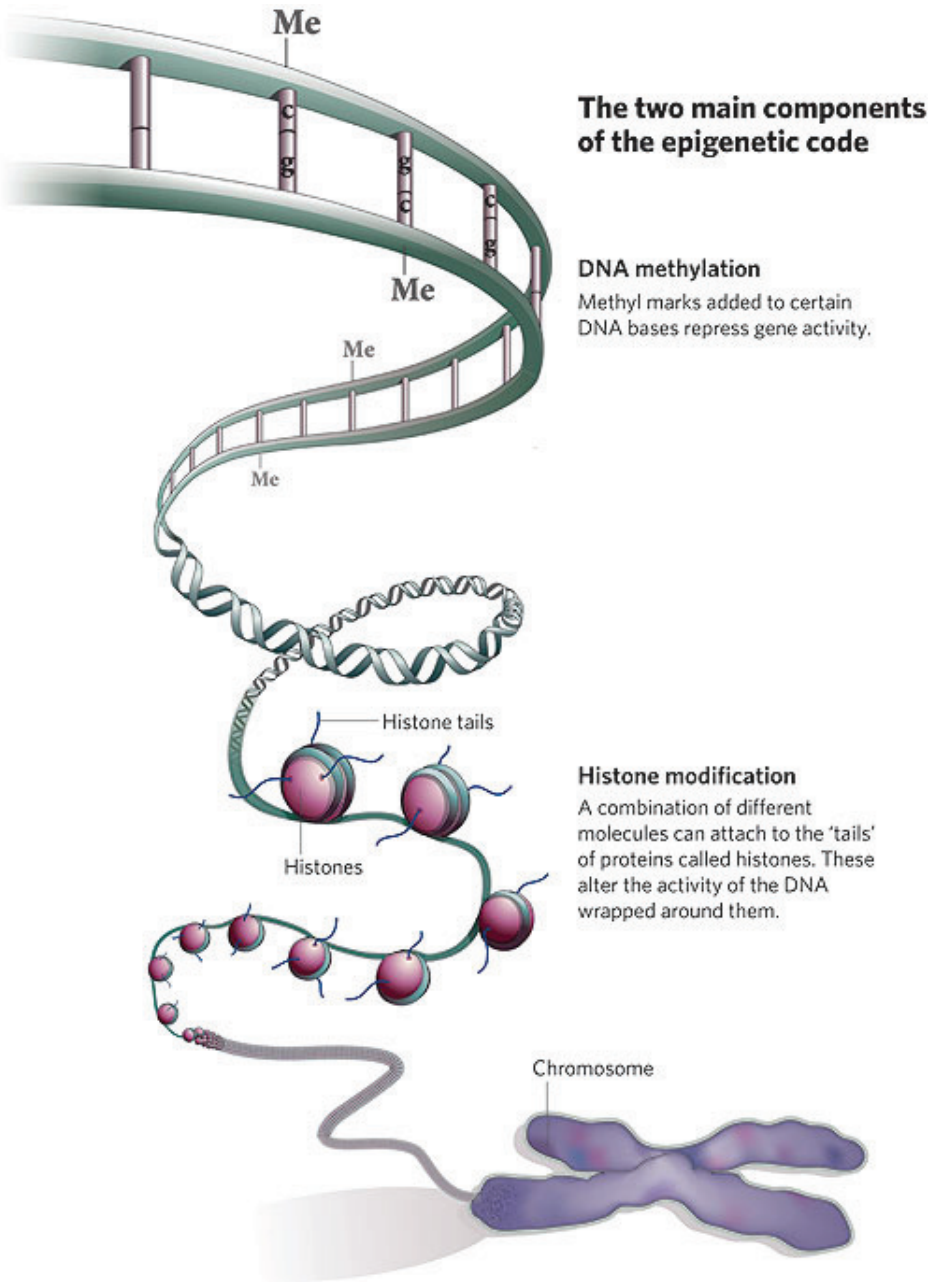

Figure 1.2. Illustration of the two main epigenetic modifications: DNA methylation and histone modifications. DNA methylation occurs when a methyl (Me) group attaches to a cytosine (C) base on the DNA molecule. Histone modifications occur when certain chemical groups attach to the tails of histone proteins. Image courtesy of Qiu (2006).

2007). Epigenetic modifications have been shown to play a role in the regulation of gene expression (Jaenisch and Bird, 2003; Vaillant and Paszkowski, 2007; Zilberman et al., 2007) and have been associated with the development of cancer (Feinberg and Vogelstein, 1983; Herman and Baylin, 2003; Feinberg and Tycko, 2004; Jones and Baylin, 2007) and other diseases (Robertson, 2005; Shames et al., 2007). Environmental factors can influence epigenetic mechanisms and one result of this interaction has been the development of promising drugs to alter epigenetic patterns in cancerous cells.

1.2.1 DNA Methylation

DNA methylation is a type of epigenetic modification that typically occurs when a methyl group (CH_3) attaches to cytosine (C) on the DNA molecule (Figure 1.2). In bacteria, adenine (A) can be methylated as well, but cytosine methylation is the most common form of DNA methylation. The addition of this chemical group to DNA does not alter the DNA sequence itself, but can have a profound impact on gene function. In 1975, two papers appeared in the literature, which suggested a potential relationship between DNA methylation and gene expression (Riggs, 1975; Holliday and Pugh, 1975). Since that time, research into this chemical mark have flourished, particularly with advances in technology in the 1990s that allow for wide-scale studies.

DNA methylation occurs in most studied organisms, with the exception of the budding yeast, *Saccharomyces cerevisiae*, and the nematode worm, *Caenorhabditis elegans*, and is limited to embryonic development in the fruit fly, *Drosophila melanogaster* (Bird, 2002; Suzuki and Bird, 2008). In mammals, DNA methylation typically occurs when a cytosine (C) is followed by a guanine (G) in the $5' - 3'$ direction of the DNA sequence. This is denoted CpG, to represent the fact that cytosine and guanine are linked together by phosphate on one of the DNA strands (Li and Bird, 2007). DNA methylation is established by a family of *de novo* DNA methyltransferase en-

zymes (DNMT3) and is maintained during DNA replication by a maintenance DNA methyltransferase (DNMT1). At some locations, known as CpG islands, the number of CpG sites given GC content in a region of certain length is higher than expected. CpG islands have been shown to occur upstream of many genes and are typically unmethylated (Law and Jacobsen, 2010).

In plants, DNA methylation occurs at both CpG locations and in CpNpG and CpNpN sequence contexts (where N is one of the nucleotide bases A, C or T). CpG and CpNpG locations are called symmetric since, due to complementary base pairing, DNA methylation can occur on both strands at those sites; whereas CpNpN sites are asymmetric since DNA methylation can only occur on one of the strands in that context. The DOMAINS REARRANGED METHYLTRANSFERASE enzyme family (DRM) serves to establish DNA methylation and is similar to the DNMT3 enzyme family in mammals. CpG methylation is mostly maintained by the DNA METHYLTRANSFERASE1 (MET1), which is similar to DNMT1 in mammals, but is supplemented by the DECREASE IN DNA METHYLATION1 (DDM1) and HISTONE DEACETYLASE6 (HDA6) genes. Maintenance of CpNpG methylation is conducted by a plant-specific methyltransferase, CHROMOMETHYLASE3 (CMT3), and CpNpN methylation is maintained by enzymes in the DRM family (Chan et al., 2005; Law and Jacobsen, 2010).

DNA methylation plays an important role in many different biological processes, including genomic imprinting, X chromosome inactivation, embryonic development, and silencing of transposable elements (Bird, 2002; Gehring and Henikoff, 2007; Slotkin and Martienssen, 2007; Kim et al., 2009; Finnegan, 2010). In plants, DNA methylation is important for genome stability and plant development (Gehring and Henikoff, 2007; Finnegan, 2010). In humans, many diseases have been linked to DNA methylation (Robertson, 2005; Shames et al., 2007). In particular, Feinberg and Vogelstein (1983) introduced the first evidence, which connected DNA methylation to the development of cancer. Since that time, researchers have shown that a global loss of DNA methylation (hypomethylation) accompanied with a targeted gain of

methylation (hypermethylation) at CpG islands in promoter regions are characteristic patterns in cancerous cells (Herman and Baylin, 2003; Feinberg and Tycko, 2004; Jones and Baylin, 2007; Shames et al., 2007).

1.2.2 Histone Modifications

In eukaryotes, chromosomes are packaged and condensed in the cell via chromatin, a combination of DNA and histone proteins. The basic unit of chromatin is the nucleosome, which contains 147 base pairs of DNA wrapped around a histone octamer. Changes to chromatin structure can regulate gene expression since transcriptional access to the DNA is limited by the chromatin packaging. One way chromatin can be altered is through post-translational modifications of the histone proteins. The four core histones present in chromatin are: H2A, H2B, H3, and H4. These histones are organized in the octamer so that their amino tails protrude from the nucleosome, allowing the possibility of the attachment of chemical groups such as acetyl, methyl, phosphate, and ubiquitin (Figure 1.2) (Kouzarides, 2007; Griffiths et al., 2008; Margueron and Reinberg, 2010). Such chemical modifications to histone proteins were identified by Allfrey et al. (1964), who postulated that histone modifications play a role in transcription. However, evidence for such a relationship was scarce until the 1990s when the first histone acetyltransferase (HAT) enzyme was identified, revealing a mechanism by which histone acetylation could arise in the cell (Brownwell et al., 1996). Since that time, understanding the role of histone modifications in gene regulation and other cellular processes has thrived (Kuo and Allis, 1998; Berger, 2002; Kouzarides and Berger, 2007).

Many relationships between histone modifications and gene activation, or repression, have been identified. For example, methylation of histone H3 at lysines K4 and K36 (denoted H3mK4 and H3mK36) activates transcription, whereas methylation of histone H3 at lysines K9 and K27 (H3mK9 and H3mK27) represses transcription (Kouzarides and Berger, 2007). Acetylation and phosphorylation have been shown

to play a role not only in gene regulation, but also in DNA repair and chromosome condensation (Kouzarides, 2007). Like DNA methylation, global changes in histone modifications have been shown to be associated with cancer. For example, loss of acetylation and methylation at certain lysines on histone H4 are a common pattern in cancerous cells (Jones and Baylin, 2007). More recently, researchers have begun investigating the relationship between histone modifications and DNA methylation, both for cancer research and, more generally, to obtain a better understanding of how these epigenetic mechanisms may work together in cellular processes (Cedar and Bergman, 2009).

1.3 Genomics and Epigenomics

An organism's complete set of genetic material (DNA) is called a genome. In the 1990s, advances in technology made it possible to move from localized genetic and epigenetic studies to genome-wide investigations. The sequencing (i.e., identifying the order of DNA base pairs) of entire genomes became a feasible task, with *Hemophilus influenza* being the first free-living organism to be sequenced (Fleischmann et al., 1995). Along with genome sequencing came the possibility of annotating the genome by identifying genes (broken down into exons and introns) and other important genomic units, such as transposable elements (Stein, 2001).

The Human Genome Project is the first major effort to sequence and identify all genes in the human genome. The first draft was released in 2001 by private (Venter et al., 2001) and public (International Human Genome Sequencing Consortium, 2001) projects, with updates in following years. Genome projects for over 1100 organisms, including many model organisms such as the fruit fly *Drosophila melanogaster*, the model plant *Arabidopsis thaliana*, and the mouse *Mus musculus*, have been completed as of September 2009 (GOLD: Genomes OnLine Database v 3.0, 2010; Liolios et al., 2010).

The massive amount of data obtained from genome projects ushered in the establishment of online biological databases to store and make information publicly available. Databases such as Genbank (National Center for Biotechnology Information, 2010; Benson et al., 2008) store nucleotide sequence data for all organisms, while the Genomes OnLine Database (GOLD: Genomes OnLine Database v 3.0, 2010; Liolios et al., 2010) and other databases collect information from multiple sources to store in a common, searchable location. Genomic annotation databases are often maintained separately for different organisms, such as The Arabidopisis Information Resource (TAIR) for *Arabidopsis thaliana*, FlyBase for *Drosophila melanogaster*, and Ensembl for human, mouse, and many other vertebrates. In addition to identifying gene locations and other genomic elements, these genomic annotation databases also give information about gene function (Stein, 2001). Information from genome projects made possible the study of functional genomics, which seeks to learn the function, expression, and interaction of gene products. One major area of functional genomics research involves the investigation of gene expression under certain conditions. The coupling of data from genome projects and the development of microarray technology has allowed the measurement of gene expression levels for all known genes in a genome (Griffiths et al., 2008).

More recently, unique designs of microarrays have also allowed the possibility of epigenomic studies, in which locations of epigenetic modifications are identified across whole genomes. This is a complex task since, unlike the DNA sequence, epigenetic modifications are variable between cell types and over time within an individual organism (Suzuki and Bird, 2008). The first genome-wide DNA methylation map was completed in the model plant *Arabidopsis thaliana* (Zhang et al., 2006). However, many epigenome projects have been initiated, such as the Human Epigenome Project, which has the goal of studying DNA methylation patterns in major human tissues, and the Alliance for the Human Epigenome and Disease (AHEAD) project, which investigates DNA methylation and specific histone modifications in specific tissues specified by the Human Epigenomic Task Force (Jones and Martienssen, 2006; AHEAD, 2008).

The importance of these epigenomic studies is to make connections between genomic annotation and locations of epigenetic modifications to identify patterns across the genome. Examining the distribution of epigenetic modifications across the genome can help researchers make connections between the fields of genetics and epigenetics for the purpose of gaining a better understanding of their effect on phenotypes.

1.4 Microarray Technology and Applications

As genome projects were starting to thrive in the late 1990s and early 2000s, making it possible to obtain whole-genome sequences and gene locations for many organisms, a technology called a microarray was developed (Schena et al., 1995) that enhanced genome-wide investigations. Microarrays require knowledge of DNA sequence information for the development of single-stranded probes, which are placed on a microarray chip and have the potential to bind to a single-stranded mRNA or DNA sample via complementary base pair binding. The first common application of microarrays was the study of mRNA transcript abundance (i.e., gene expression level) to determine which genes are active in making proteins in a given sample (Schena et al., 1995; Lashkari et al., 1997). Thousands of genes (often all known genes in an organism's genome) can be represented on one microarray chip and investigated in a single experiment. From gene expression microarray technology, a different type of microarray, known as a tiling array, evolved with probes that cover the whole genome not just gene regions. This whole-genome coverage of tiling arrays made it possible to use microarray technology to study epigenetic modifications, which can occur anywhere in the genome (in genes and intergenic regions) (Mockler and Ecker, 2005).

1.4.1 Gene Expression Microarrays and Differential Expression Studies

Gene expression microarrays are designed to measure mRNA transcription levels of thousands of annotated genes on a single array. A common goal of gene expression

microarray experiments is differential expression analysis. In such studies, transcription levels of all genes represented on the array are measured and compared between different conditions of interest (e.g., treatment vs. control) to obtain a set of genes that exhibit statistically significant different expression levels between conditions. Such studies can help determine which genes are important for different biological processes and diseases.

While different gene expression arrays have been developed, a commonly used array platform for gene expression studies are oligonucleotide arrays commercially produced by Affymetrix® (Lockhart et al., 1996; Lipshutz et al., 1999). For Affymetrix® gene expression arrays (GeneChips), 25 base long DNA probes from a reference genome are selected from annotated genes of a specific organism and placed as targets on the microarray. An mRNA sample collected from that organism is then amplified, labeled with a fluorescent dye, and hybridized to the array through complementary base pair binding. Two types of probes represent each sequence, a perfect match (PM) probe which matches all 25 bases of the reference sequence and a mismatch (MM) probe, designed to measure non-specific binding, which differs at the 13^{th} base. After hybridization, mRNA transcription levels for each probe are measured in the form of a quantitative intensity reading and can be used to indicate which genes are active in making proteins in that sample (Figure 1.3). Note that one dye is used in Affymetrix GeneChips®, so one array per mRNA sample is required.

The selection of probes is an important component for the success of microarray technology. Since the goal is to measure transcription levels of known genes, it is imperative that probes on the array correspond to exons of genes where transcription, and thus hybridization of the mRNA sample to probes, is expected to occur. Typically, on an Affymetrix® gene expression array, genes are represented by 11-20 probes (called a probe set) that cover exons of genes and are chosen for their optimal hybridization quality (Affymetrix® Technical Note, 2007). Figure 1.4 gives an example of probes covering a genomic region for a typical Affymetrix® gene expression microarray.

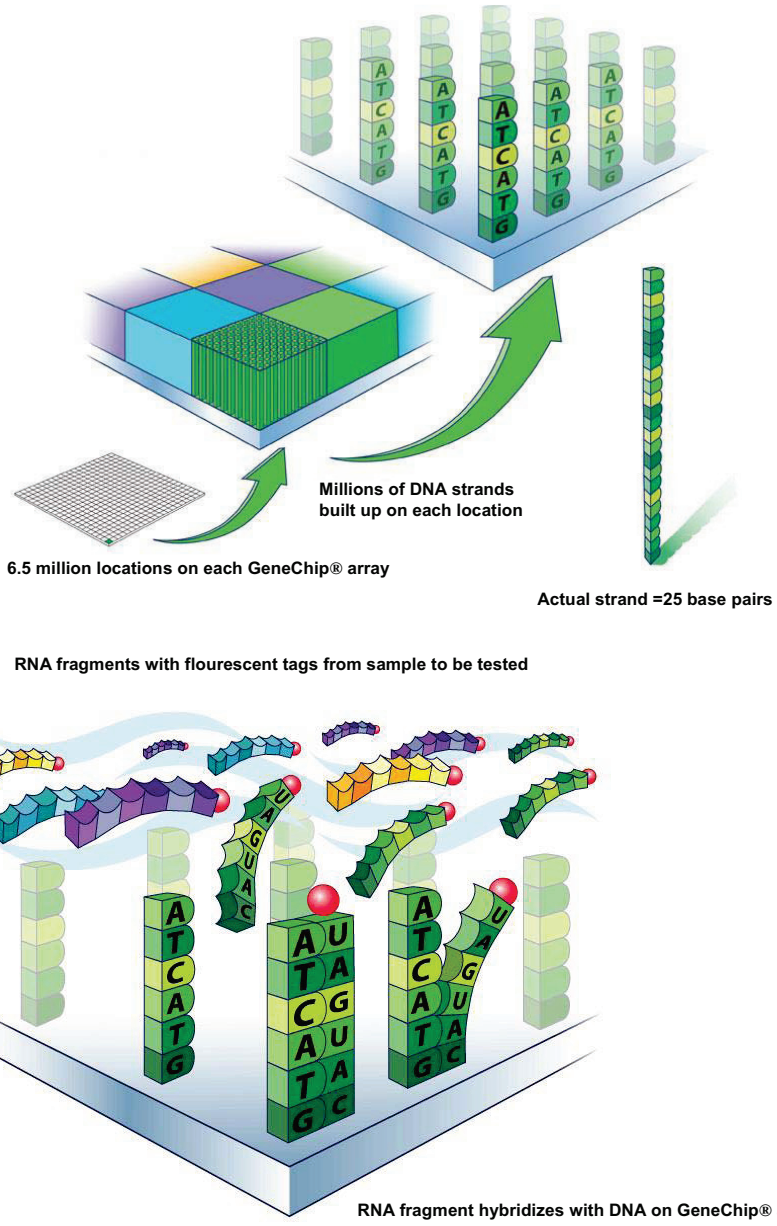

Figure 1.3. The top figure shows one feature on an Affymetrix GeneChip® microarray. Each feature contains millions of copies of a DNA probe that is 25 bases in length. The bottom figure shows hybridization of an mRNA sample to the array. Fluorescent labeling allows the calculation of a numerical intensity reading which represents the amount of transcription (gene expression) taking place per gene. Image courtesy of Affymetrix® Image Library (2009), www.affymetrix.com.

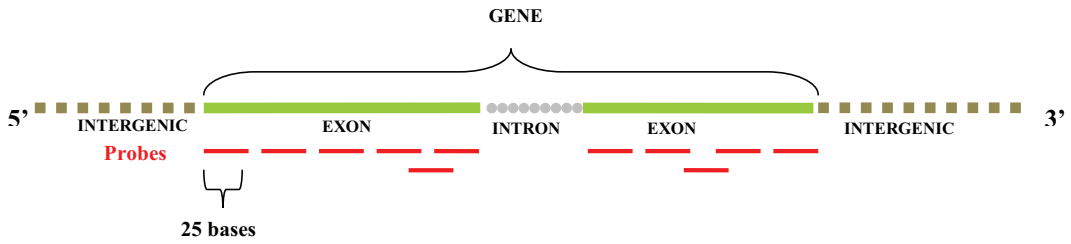

Figure 1.4. Example of probes covering a genomic region on an Affymetrix® gene expression microarray. The 25 base long probes cover exons of genes and can be overlapping. Each gene is typically represented by 11-20 probes.

One alternative to Affymetrix® arrays for studying gene expression is custom-designed spotted cDNA microarrays (Schena et al., 1995). These arrays are similar to Affymetrix® arrays in that they rely on target sequences placed on the array and complementary base pairing for hybridization. However, instead of representing each gene by a set of short oligonucleotide probes, the probes on these arrays represent complementary DNA (cDNA) of whole genes or expressed sequence tags (ESTs) that represent the gene. Sequences for the probes can be isolated and amplified via polymerase chain reaction (PCR), then purified and printed on the microarray chip by a robot. These probes are typically longer than the Affymetrix® oligonucleotide probes, as they depend on the length of the gene or EST. Typically, for spotted cDNA arrays, an mRNA sample is collected, converted to cDNA, and then labeled with one of two fluorescent dyes (Cy3 - green or Cy5 - red). A second mRNA sample (e.g., one treatment and one control sample) undergoes the same preparation, but is labeled with the other dye. The samples are mixed and hybridized to the same array. Intensity readings for each dye give the mRNA transcript levels. Genes with a high red or green reading mean that gene was expressed more in one of the samples; whereas genes with yellow spots indicate the expression level was similar for the two samples and black spots are genes which were not expressed in either sample (Schena et al., 1995; Duggan et al., 1999). These arrays are often useful for smaller-scale studies in

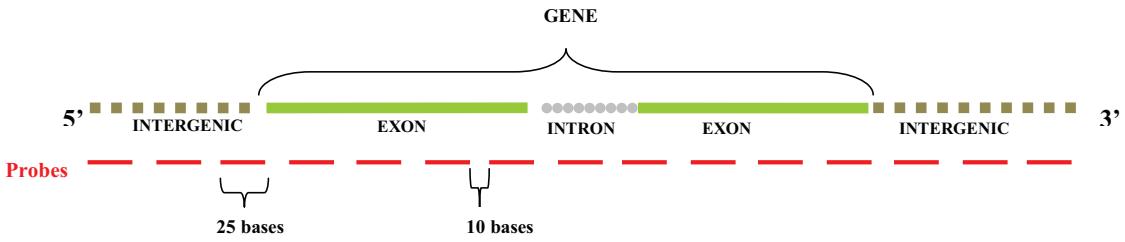

Figure 1.5. Example of probes covering a genomic region on an Affymetrix® tiling array. Probes are systematically placed from one end of the region to the other without regard for genomic annotation. As a result, exons and introns of genes, as well as intergenic regions, are covered by probes. In this example, probes are 25 bases in length with an average gap of 10 bases between probes.

which scientists are interested in specific genes or regions of the genome, since they can select which genomic sequences to place as targets on the array.

1.4.2 Tiling Arrays and Applications

Tiling microarrays work in a similar fashion as gene expression arrays, in that DNA probes are selected from a reference genome and placed as targets on the array for hybridization with a sample of genetic material (mRNA or DNA). However, the key difference between the gene expression microarray and the tiling microarray is in the probe selection process. Tiling arrays are designed to cover entire genomic regions (e.g., chromosomes or whole genomes) by systematically selecting probes from one end of the region to the other without regard to genome annotation. These probes, often called tiles for tiling arrays, are not specifically designed to optimize the study of gene expression, but rather to provide a dense, unbiased coverage of the genomic region. As with gene expression arrays, both Affymetrix® and spotted cDNA arrays are available. Figure 1.5 shows an example of probes covering a genomic region for a typical Affymetrix® tiling array. Affymetrix® tiling array probes are typically 25

bases in length, but the spacing or overlap between these probes differs by organism. Because they cover entire regions rather than just exons of genes, tiling arrays can be used for a variety of applications. They are particularly well-suited for studying epigenetic modifications that may occur anywhere in the genome (Mockler and Ecker, 2005).

Differential Expression Studies

While gene expression arrays have been used to study differential gene expression for many years, relatively few studies have focused on using tiling arrays for differential expression analysis (Ghosh et al., 2007; Naouar et al., 2009; Zeller et al., 2009). Many studies have used tiling arrays to study transcription, however their focus has been on transcript mapping, where regions of transcription are identified through statistical models (Kapranov et al., 2002; Yamada et al., 2003; Bertone et al., 2004; Kampa et al., 2004; Schadt et al., 2004; Huber et al., 2006). Tiling arrays are well-suited for this purpose since their dense coverage can lead to the identification of novel transcripts and can improve genome annotation. While new regions of transcription or transcript variants will continue to be found, making use of the current genome annotation to obtain differential expression results for known genes is a common practical need for researchers. This type of analysis can be achieved with both gene expression and tiling arrays. However, there is a large difference in genomic coverage between these arrays since tiling arrays cover the whole genome and gene expression arrays cover exons of genes. While this extra information on the tiling array is important for epigenetic tiling array applications, only the probes in exons of genes are of primary interest when the goal is studying differential expression of known genes. Thus, knowing where probes are located with respect to genome annotation and keeping only biologically relevant probes for a statistical analysis is an essential step for differential expression studies with tiling arrays.

DNA Methylation Profiling

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 dense, unbiased genomic coverage and they 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). In order to detect DNA methylation status of individual probes represented on the array, genomic DNA samples must be prepared in a certain way before hybridization to the tiling array (e.g., Figure 1.6). Genomic DNA collected from one individual is split into two samples and sheared into similar sizes. In one of these samples, a treatment such as bisulfite conversion, digestion with a methylation sensitive restriction enzyme such as McrBC, 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 retained. Double-stranded DNA from both the treated sample (which will consist of only methylated or unmethylated DNA) and the untreated sample are separated to single-stranded DNA and hybridized to tiling arrays (Weber et al., 2005; Keshet et al., 2006; Schumacher et al., 2006; Beck and Rakyan, 2008; Estecio and Issa, 2009). Hybridization intensities between the treated and untreated samples are compared for each probe via a statistical model to estimate whether the probe is methylated or not. Typically, these results are then visually connected to genomic annotation to gain an understanding of the distribution of DNA methylation across the genomic region.

Histone Modification Studies

Tiling arrays can also be used to study histone modifications through a technique called ChIP-chip (chromatin immunoprecipitation combined with microarray chips).

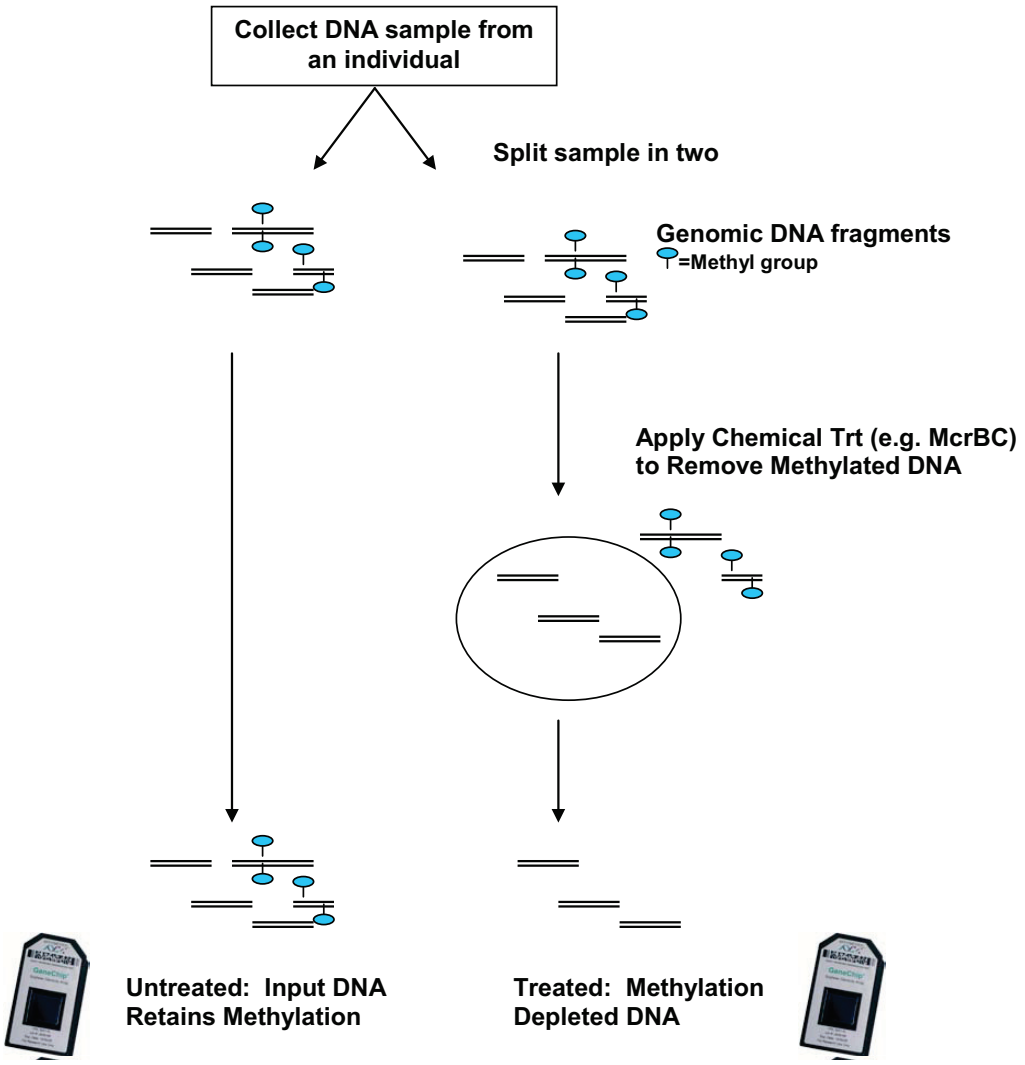

Figure 1.6. Illustration of DNA sample preparation technique for DNA methylation profiling studies with tiling arrays. In this example, DNA from one individual is split in two and sheared into similar sizes. Treatment by digestion with a methylation restriction enzyme (e.g. McrBC) is employed in one sample to remove methylated DNA. No treatment is applied to the other sample, which retains both methylated and unmethylated DNA. Single-stranded DNA from both of these samples is then hybridized to tiling arrays. Affymetrix® microarray images courtesy of Affymetrix® Image Library (2009)

Chromatin immunoprecipitation investigates the interaction between DNA and proteins. During the immunoprecipitation process, it is possible to isolate DNA that is linked to a specific protein (e.g., DNA that is wrapped around a histone protein with a certain chemical attachment) using a protein-specific antibody. Similar to the DNA profiling sample preparation, a control (untreated) sample and the immunoprecipitated sample are then hybridized to tiling arrays. Comparing hybridization intensities between these two samples via a statistical model gives estimates for which probes are enriched or depleted of that protein (Buck and Lieb, 2004).

1.5 Genomic Annotation and Tiling Array Experiments

Tiling arrays have enabled researchers to investigate both gene expression and epigenetic modifications on a genome-wide scale. Although probe selection without regard to genome annotation is an essential design component of tiling arrays, the knowledge of which probes belong to which genomic regions can provide valuable information in statistical analysis. In differential expression studies, for example, annotation information can be used to select biologically relevant probes by retaining probes that cover exons and filtering out probes that correspond to introns and intergenic regions. In DNA methylation profiling studies, neighboring probes are likely to be correlated and often DNA methylation patterns based on genomic annotation are identified after a statistical analysis is complete. Incorporating information from both neighboring probes and genomic annotation into statistical methods may help improve prediction of probe DNA methylation status, and is the focus of this dissertation. Specifically, the use of genomic annotation information in statistical methods for both differential expression analysis and DNA methylation profiling with tiling arrays is explored.

2. GENOMIC ANNOTATION OF TILING ARRAYS

A key feature offered by tiling arrays is unbiased, wide-scale genomic coverage. This feature allows for the study of epigenetic modifications and other biological phenomena that were not possible to investigate with gene expression arrays. Although selecting probes without regard to genome annotation provides this unbiased coverage, it is often of interest to connect results from tiling array experiments to genome annotation for interpretation after a statistical analysis is complete. The genomic position and sequence of probes is often provided in a data file containing information about the array platform. However, information about which probes correspond to exons, introns, or intergenic regions is not always given (Figure 2.1).

Online genome annotation databases can be employed to obtain the location of exon and intron regions of known genes for specific organisms. These data can be linked to probe position information to determine which genomic element each probe represents (Figure 2.2). Connecting probes to genomic annotation not only aids in interpretation of tiling array data results, but also enables the possibility of incorporating genomic annotation information into statistical methods. This work investi-

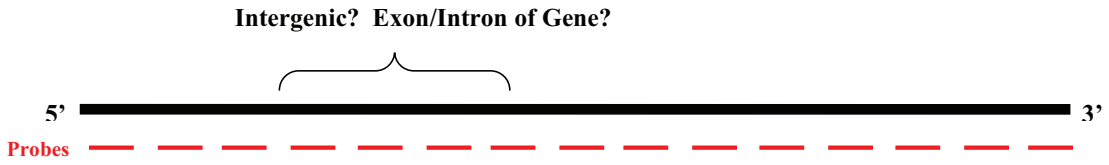

Figure 2.1. Example of a genomic region without genome annotation information. Genomic locations of tiling array probes are known, but information about whether probes correspond to exons, introns, or intergenic regions is unknown.

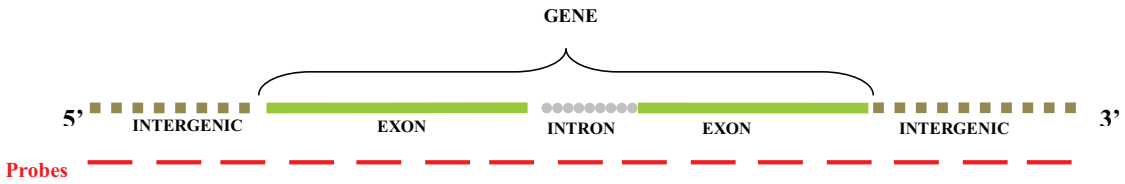

Figure 2.2. Example of a genomic region after connecting tiling array probes to genome annotation information. Both the genomic locations of tiling array probes and the genomic elements the probes represent is known. In this example, information about which probes correspond to exons, introns, or intergenic regions is given.

gates the benefits of utilizing genome annotation in differential expression and DNA methylation tiling array data analysis.

2.1 Genomic Annotation and Differential Expression Analysis

Recall that differential expression studies aim to identify a set of genes with statistically significant different expression levels between conditions of interest (e.g., treatment vs. control). Ideally, statistical testing is performed on a gene level for all annotated genes represented on the array. However, without genomic annotation information, it is unknown which tiling array probes correspond to exons of genes (Figure 2.1), where transcript accumulation is typically expected to occur. Testing for differential expression between conditions is therefore limited to testing each of the probes on the tiling array individually.

Probe level testing is problematic for many reasons. Probe level tests make it challenging for scientists to obtain practical results that can be interpreted on a gene level. For instance, if some probes are identified as significantly different for a particular gene and other probes are not, it is difficult to draw conclusions about differential expression for that gene. Furthermore, whole-genome tiling arrays often contain millions of probes. Testing on a probe level greatly increases the well known

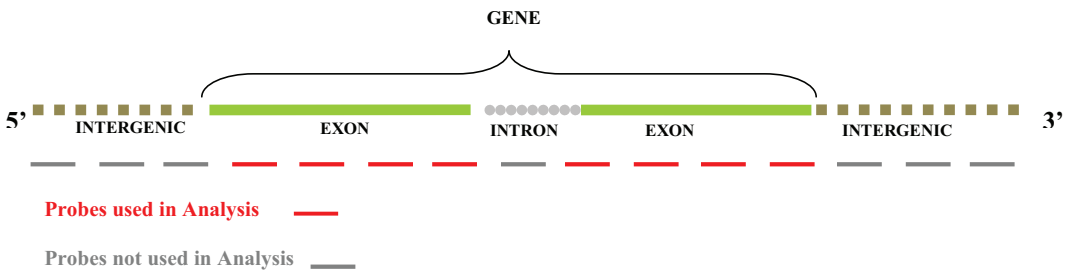

Figure 2.3. Example of selecting biologically relevant tiling array probes for differential expression analysis using genome annotation information. Red probes correspond to exons of genes and are retained for differential expression analysis; whereas light grey probes cover introns and intergenic regions and are filtered out before analysis. The set of red probes for this gene are considered to be a probe set.

statistical challenge of multiple testing since the magnitude of the number of tests increases from the thousands (number of genes) to the millions (number of probes). Many of these probes are not even of primary interest since they correspond to introns or intergenic regions. These issues can be resolved by connecting tiling array probes to genomic annotation and using biologically relevant probes for differential expression analysis. Selection is made by retaining probes covering exons and filtering out probes corresponding to introns and intergenic regions. The resulting data have the same format as gene expression arrays with each gene containing multiple probes in exons of genes to form a probe set (Figure 2.3).

2.2 Genomic Annotation and DNA Methylation Profiling

In DNA methylation profiling studies with tiling arrays, statistical methods are employed to determine whether each probe is methylated or not. Since DNA methylation can occur anywhere in the genome at certain cytosine sites, all of the probes are biologically relevant and methylation status of each individual probe estimated. It is common in such studies to connect estimated probe methylation status to genomic

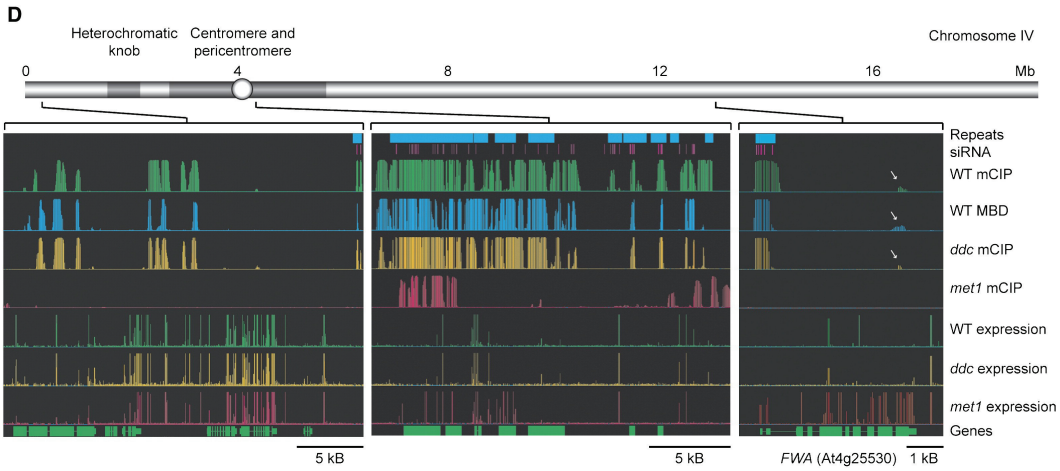

Figure 2.4. Example of linking genomic annotation to DNA methylation analysis results (Zhang et al., 2006). A euchromatic region (left), a heterochromatic region (middle), and the FWA (late flowering mutant) locus of chromosome four of *Arabidopsis thaliana* are shown. The top two tracks give locations of repetitive elements and siRNAs. The seven tracks below that give DNA methylation and gene expression results for three different types of *Arabidopsis* (WT, *ddc*, and *met1*). The bottom track shows where genes are located. Results are compared visually and summarized by genomic element across the region to identify DNA methylation patterns. Image courtesy of Zhang et al. (2006).

annotation after the analysis is complete (Figure 2.4). Patterns of DNA methylation are then investigated according to different types of genomic elements.

Different organisms show different overall methylation patterns. Mammals often exhibit a global pattern, where DNA methylation is found at most CpG sites throughout the genome, with the exception of CpG islands which are typically unmethylated. 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 2.5) (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

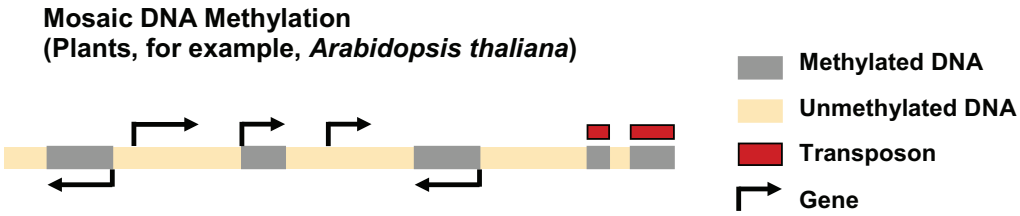

Figure 2.5. Example of mosaic DNA methylation. As defined by Suzuki and Bird (2008), mosaic methylation occurs when densely methylated regions (grey) are interspersed with unmethylated or less densely methylated regions (yellow). In this example genes (arrows) are either heavily methylated or completely unmethylated, and transposons (red) are methylated.

20% of the genome exhibiting dense methylation. These studies suggest that dense DNA methylation typically occurs in transposons, which are targeted for methylation by an RNA-mediated defense mechanism, and inactive heterochromatin, including the darkly stained region on chromosome 4 known as the heterochromatic knob. 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). The pattern of longer regions of DNA methylation in certain genomic regions (e.g., transposons, gene bodies) interspersed with unmethylated or less densely methylated regions, suggests that incorporating genome annotation information into statistical methods used to detect locations of DNA methylation in *Arabidopsis* (and other organisms with similar patterns) may be valuable.

2.3 *Arabidopsis thaliana* Tiling Arrays

Arabidopsis thaliana is a small mustard plant that serves as the model organism for plants. Due to its short life cycle (six weeks) and small size, it is an ideal organism for lab experiments (Meinke et al., 1998). *Arabidopsis thaliana* also has a small genome

size with five chromosomes and around 30,000 genes (The Arabidopsis Information Resource (TAIR), 2008).

Arabidopsis thaliana research has been at the forefront of the fields of genomics and epigenomics. It was the first plant (and third multi-cellular organism) to be fully sequenced (The Arabidopsis Genome Initiative, 2000), and the first gene expression microarray was based upon *Arabidopsis thaliana* DNA sequences (Schena et al., 1995). Genome annotation information is stored at the The Arabidopsis Information Resource (TAIR) (2008) website (`www.arabidopsis.org`), with updates to gene structures made 1-2 times per year. In epigenomics research, it was the first organism for which a genome-wide DNA methylation study was conducted (Zhang et al., 2006). In this work, two different tiling arrays (a custom-designed cDNA array and an Affymetrix® array) developed for*Arabidopsis thaliana* are used as motivating examples to show how genomic annotation information can be incorporated into statistical methods for differential expression and DNA methylation studies. A summary of the design and genomic annotation of these two arrays is given in this chapter.

2.3.1 Custom-designed Chromosome 4 Tiling Array

Lippman et al. (2004) custom-designed a cDNA spotted tiling array to investigate gene expression, DNA methylation, and certain histone modifications of a heterochromatic knob (*hk4S*) on chromosome four of *Arabidopsis thaliana*. Heterochromatin is highly condensed chromatin that is less accessible for transcription than its counterpart, euchromatin. Heterochromatin is known to contain repetitive DNA and transposons, and heterochromatic DNA is often heavily methylated (Martienssen and Colot, 2001). Although heterochromatin is typically found at centromeres and telomeres, sometimes regions of heterochromatin are present in other parts of the chromosome called "knobs" (Grewal and Jia, 2007). In *Arabidopsis*, one of these heterochromatic knobs is found on the short arm of chromosome 4 (*hk4S*).

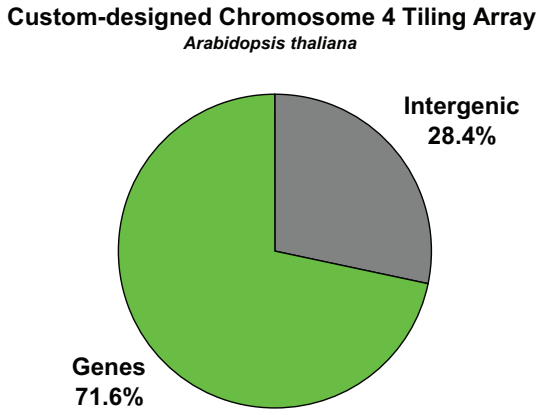

Figure 2.6. Percentage of probes mapping to genes and intergenic regions on the Lippman et al. (2004) custom-designed tiling array for the chromosome 4 heterochromatic knob (*hk4S*) in *Arabidopsis thaliana*.

The Lippman et al. (2004) tiling array contains 1722 unique probes. Of these, 1407 probes cover a 1.5-megabase (Mb) region centered on *hk4S* with representation of a small euchromatic region on both sides of the knob. Each probe is replicated two to four times on the array. Recall that spotted cDNA array probes are often longer than Affymetrix® probes and, on this array, the average probe length is 995 bases with an average gap of 56 bases between probes (although many probes overlap). Connecting the probes to genomic annotation reveals that a majority of the probes (71.6%) lie in gene regions (Figure 2.6), with an average of three probes representing each gene. A set of 680 probes from the euchromatic region can be used as controls, as they are less likely to be methylated than probes in the heterochromatic knob.

2.3.2 Affymetrix® Whole Genome Tiling Array

The *Arabidopsis* Affymetrix® Tiling Array 1.0F/R is a whole genome tiling array that has been used in many different types of studies since its release in 2006 (e.g., Zhang et al. (2008); Hazen et al. (2009); Naouar et al. (2009); Zeller et al. (2009))

Table 2.1
Number of tiling array perfect match (PM) probes per chromosome on the Affymetrix® whole genome tiling array.

Chromosome	Number of PM probes
1	779,303
2	507,749
3	600,472
4	469,660
5	682,807
TOTAL	3,039,991

and was the technology used for the first genome-wide DNA methylation profiling study (Zhang et al., 2006). The *Arabidopsis* Affymetrix® Tiling Array 1.0F/R covers the entire *Arabidopsis* genome by placing 25 base probes along non-repetitive regions with an average gap (and no overlap) of 10 bases between probes. Probe sequences are based on version 5 of The Institute for Genome Research (TIGR) *Arabidopsis* database which was completed in 2004 (Affymetrix® Package Insert, 2006). Note that TIGR hosted the online genome annotation database for *Arabidopsis* before it begun being maintained by TAIR, starting in 2005.

There are 3,039,991 tiling array probes covering the five *Arabidopsis* chromosomes. Table 2.1 gives the number of perfect match (PM) probes on each of the five chromosomes. Approximately 0.4% ($\sim$ 12,000) of these probes are repetitive since their sequence is represented at more than one genomic location. However, none of the probes are repeated on their own chromosome.

Tiling array probes are mapped to their genomic annotation by using data from the The Arabidopsis Information Resource (TAIR) (2008) website (`www.arabidopsis.org`). Each probe is mapped to an exon, intron, or intergenic region of the TAIR8 genome. For a probe to be considered part of an exon, it must overlap with at least one base of an exon. Since genes can have alternative isoforms, only TAIR8

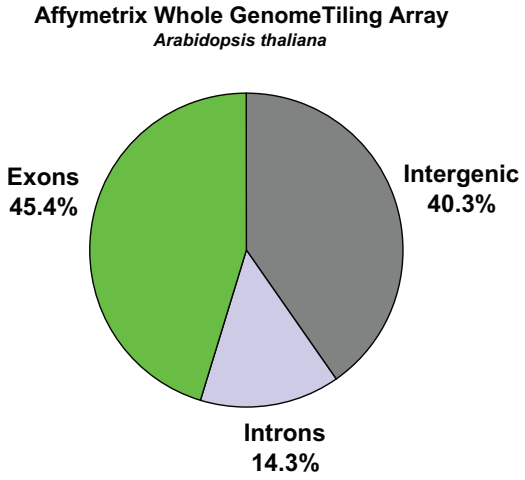

Figure 2.7. Percentage of probes mapping to exons, introns, and intergenic regions on the Affymetrix® whole genome tiling array for *Arabidopsis thaliana*.

representative gene models are used in the mapping. A large percentage of probes (45.4%) correspond to exons of genes, while 14.3% correspond to introns and 40.3% are in intergenic regions (Figure 2.7). Since some genes overlap, there are a small number of probes (0.4%) that correspond to more than one gene. There are 31,391 genes represented by probes in exons, averaging 44 probes per gene and covering 95% of TAIR8 genes.

3. USING GENOMIC ANNOTATION FOR DIFFERENTIAL EXPRESSION ANALYSIS WITH TILING ARRAYS

Gene expression can be studied using both gene expression and tiling microarrays. Understanding the design differences between these arrays is essential for developing statistical methods to test for differential expression between conditions. Recall that gene expression arrays cover exons of genes, while tiling arrays cover whole genomic regions without regard to annotation. As described in Chapter 2, tiling array probes can be connected to their genomic annotation to identify biologically relevant probes for a differential expression analysis. This is accomplished by filtering out probes inside introns and intergenic regions and retaining probes covering exons. This initial bioinformatic step yields tiling array data in the same form as gene expression array data, with multiple probes corresponding to a probe set for each gene.

Statistical issues for differential expression analysis using gene expression arrays have been thoroughly investigated (e.g., Kerr et al. (2000); Wolfinger et al. (2001); Bolstad et al. (2003); Irizarry et al. (2003); Smyth (2004)) and many analysis methods are available through statistical packages such as R/Bioconductor (R Core Development Team, 2009; Gentleman et al., 2004). With tiling and gene expression array data in the same form, statistical methodology developed for gene expression arrays can also be employed to conduct gene level tests for tiling arrays. Here, real data from an *Arabidopsis thaliana* experiment, where the same mRNA samples are hybridized to both Affymetrix® gene expression and tiling arrays, are used to compare the two array types and demonstrate the application of genomic annotation to tiling array data. Note that the work in this chapter has been described previously in Olbricht et al. (2009) and parts of the chapter are taken from that text[1].

[1]The contents of this chapter are adapted from a previous publication (Olbricht et al., 2009). Some sections are taken verbatim from Olbricht et al. (2009) while others have been revised or paraphrased.

3.1 Comparison of Affymetrix® Gene Expression and Tiling Arrays

A comparison of Affymetrix® gene expression (ATH1 array) and tiling array designs for *Arabidopsis thaliana* highlights some potential advantages and disadvantages of each array type for differential expression analysis. Note that both arrays use 25 base probes, however there is a magnitude of difference in genomic coverage on the two arrays. On the ATH1 array, there are 251,078 probes which correspond to 22,810 probe sets. Some probe sets correspond to more than one gene resulting in coverage of 23,087 (70%) of TAIR8 genes. In comparison, recall that the Affymetrix® *Arabidopsis* tiling array contains 3,039,991 probes which cover 31,391 (95%) of TAIR8 genes. There are 22,850 TAIR8 genes that are present on both arrays.

One potential advantage of tiling arrays is that they provide more genomic coverage per gene than gene expression arrays. Affymetrix® *Arabidopsis* tiling arrays average 44 probes per gene compared to an average of 11 probes per gene for ATH1 arrays. However, ATH1 arrays have a potential advantage in the probe selection process, since probes on that array are selected for optimal hybridization quality/ability for gene expression; whereas tiling array probes are not designed specifically to optimize the study of gene expression.

In addition to providing more dense coverage, tiling array annotation can be based on the most current genome version. Although the sequences of the Affymetrix® *Arabidopsis* tiling array probes are based on the TIGRv5 genome version, it is possible (due to their dense coverage) that as more up-to-date annotation becomes available tiling arrays will have probes covering newly discovered genes. Gene expression arrays, however, must be based on the annotation available when the array is made. If new genes are discovered, a new array platform must be made to investigate them. For *Arabidopsis*, the ATH1 array is based on known genes available as of December 2001 in The Institute for Genome Research (TIGR) database (Affymetrix® Data Sheet, 2004), and any information from more recent genome versions is not incorporated in the design. This accounts for the difference between the 23,087 genes represented on

the ATH1 array and the 31,391 genes represented on the Affymetrix® *Arabidopsis* tiling array. This said, there is one convenient aspect of the ATH1 array that is not available for the Affymetrix® tiling array, namely the availability of a file from Affymetrix® that connects probes on the array to their corresponding probe sets. The TAIR website (`www.arabidopsis.org`) provides additional data that links probe sets to genes in the current genome annotation version, but this annotation is not readily provided by Affymetrix® for tiling arrays.

3.2 Statistical Methods

A common goal of a differential expression study based on annotated genes is to determine for each gene whether or not there is a significant difference in expression levels between conditions (e.g., treatment vs. control). Whereas this is a common application of gene expression arrays, it is only with the availability of genomic annotation for tiling arrays that it is possible to conduct gene level tests for differential expression. After filtering out introns and intergenic tiling array probes, Affymetrix® gene expression and tiling arrays have data for each gene in the form of probe sets and thus the same statistical model can be applied to both array types. This results in 31,391 and 22,810 probe set level tests for the tiling and ATH1 array, respectively.

Drawing from the literature for gene expression arrays, the following differential expression analysis can be conducted for either array type. Note that these methods are only one possibility of the many options available for statistically testing differential expression. First, the perfect match (PM) intensities are pre-processed using a robust multi-array analysis (RMA) background correction (Irizarry et al., 2003) and a quantile normalization (Bolstad et al., 2003), setting the distribution of all arrays to be equal. An analysis of variance (ANOVA) model is performed to detect probe sets which are differentially expressed between two treatment groups using the natural log of the background corrected, normalized PM intensities as the gene expression level. The ANOVA model (3.1) fit for each probe set is similar to the two-step ap-

proach employed by Wolfinger et al. (2001) and extended by Chu et al. (2002) for Affymetrix® arrays is:

$$y_{ijk} = \mu + T_i + P_j + (TP)_{ij} + \epsilon_{ijk} \tag{3.1}$$

where $i = 1, 2; j = 1, ..., p; k = 1, ..., n$; and y_{ijk} is the gene expression level for the k^{th} replicate of probe P_j under treatment T_i, μ is the average gene expression level over all probes, treatments, and replicates, T and P are the treatment and probe main effects, TP is the interaction between treatment and probe, and ϵ_{ijk} are independent errors that are normally distributed with mean 0 and variance σ^2.

To determine if there are statistically significant differences in expression between two treatment groups, the following hypotheses are tested for each probe set:

$$H_o : T_1 - T_2 = 0 \qquad \text{vs.} \qquad H_a : T_1 - T_2 \neq 0. \tag{3.2}$$

The test statistic is:

$$\frac{\overline{y}_{1..} - \overline{y}_{2..}}{\sqrt{\frac{2*MSE}{np}}} \sim t_{2p(n-1)} \qquad \text{under} \quad H_o \tag{3.3}$$

where the mean squared error (MSE) and the number of probes (p) from model (3.1) will differ for each probe set.

Testing for differential expression at each probe set results in thousands of hypotheses tests that are conducted simultaneously in a single experiment. For a single test, the probability of a Type I error (i.e., a false positive declaring a gene is differentially expressed when it truly is not) is controlled by setting the significance level (α). However, when all tests are considered together, the chance of at least one false positive increases with the number of independent tests being performed. This issue is known as the multiple testing problem and several procedures have been developed to control different variations of the Type I error rate for a set of simultaneous tests while also considering the power of the tests. In this work, two approaches are employed to adjust for multiple testing. The Holm adjustment controls the familywise error rate, which is the probability of making at least one false discovery among the probe set level tests (Holm, 1979). Benjamini and Hochberg's method controls the

false discovery rate (FDR), which bounds the expected rate of false discoveries (Benjamini and Hochberg, 1995). Holm's procedure is more conservative than that of the FDR approach.

3.3 Analysis of *Arabidopsis thaliana* Data

Data from an *Arabidopsis thaliana* study are used to demonstrate the application of tiling arrays for studying differential expression, as well as to compare tiling and gene expression array results. In this study, an overexpressing *Arabidopsis* line of the MTF gene (Myb4) is compared to wild-type Columbia (Col-0) *Arabidopsis*. The MTF gene is a myb transcription factor that, when mutated, can increase the plant's susceptibility to allowing the transfer of foreign DNA from Agrobacterium to the plant (Gelvin, 2003). Gene expression is measured by hybridizing the same mRNA samples from the root tissue of both *Arabidopsis* sample types to both Affymetrix® ATH1 and tiling arrays. Three biological replicates of each of the two sample types are used, yielding a total of 6 arrays of each type. The goal is to identify differentially expressed genes between Col-0 and Myb4.

Since the same biological samples are hybridized to both array types, discrepancies in results should be due to technological differences between the arrays or other experimental factors, such as RNA degradation, rather than due to biological differences in the samples. The same ANOVA model (3.1) is applied to data from both array types where the treatment effect is the sample type (Col-0, Myb4). The hypotheses for differential expression between the two sample types (3.2) are tested via test statistic (3.3) for each probe set, where n=3 and p can differ for each probe set.

The FDR and Holm's procedures identified 4228 and 660 significant differentially expressed probe sets, respectively, on the ATH1 array at α=0.05 (Figure 3.1, top). On the tiling array, FDR/Holm's identified 2285 and 510 probe sets that exhibited significant differential expression (Figure 3.1, bottom). Figure 3.1 shows the average log fold change of each probe set for both arrays. A positive log fold change indicates

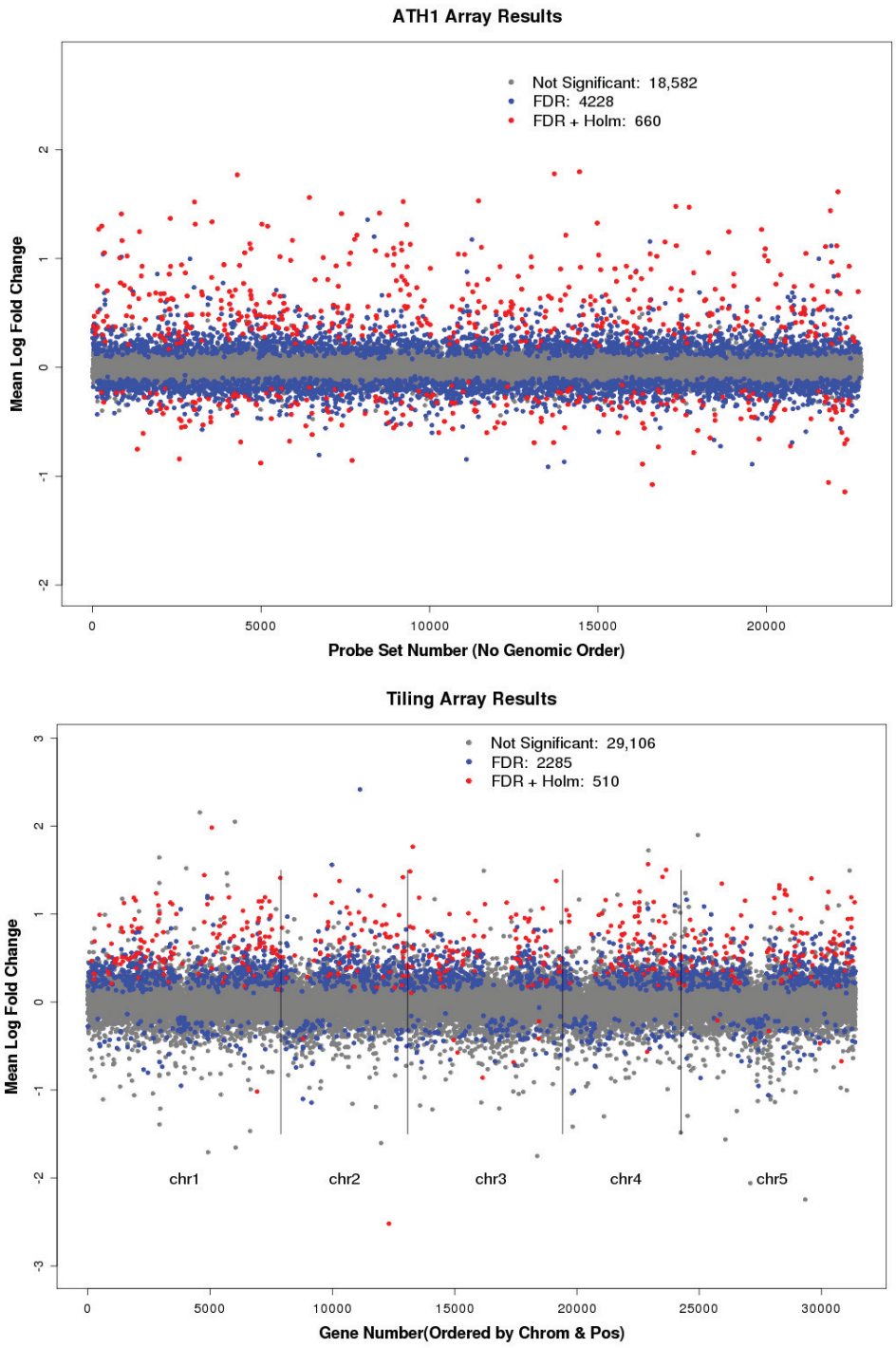

Figure 3.1. The top figure shows the mean log fold change vs. probe set number for the ATH1 array. The bottom figure shows the mean log fold change vs. gene number for the tiling array. The gene numbers are ordered by chromosomal position for the tiling array, while probe set numbers for the ATH1 array do not correspond to genomic order since one probe set can represent more than one gene. For both graphs, probe sets that are not significant are grey, probe sets significant with FDR only are in blue, and probe sets significant with both FDR and Holm's are in red. The numbers in the legend correspond to the number of probe sets in each of the three groups.

up-regulation in Col-0 compared to Myb4 and a negative log fold change is indicative of down-regulation in the Col-0 sample. Note that probe sets in the ATH1 graph (Figure 3.1, top) are not ordered since some probe sets correspond to more than one gene; whereas each probe set on the tiling array corresponds to one gene and can be ordered by the gene's position on the chromosome (Figure 3.1, bottom). The tiling array identified almost half as many differentially expressed probe sets using the FDR procedure as the ATH1 array, with the majority of significant probe sets demonstrating up-regulation and a loss of down-regulation compared to the ATH1 results.

To compare the results of differential expression in terms of genes rather than probe sets, the 22,850 genes that are present on both arrays are investigated. While many of the same significant differentially expressed genes are identified using both array types (1046 with FDR), there are also many genes uniquely identified as significant on one array but not the other (Figure 3.2). To investigate the discrepancies in the array results, one option is to look at the similarity of the mean log fold change for genes represented on both arrays. If both arrays are performing similarly at the gene level, the average log fold change for a particular gene will be similar on both arrays (since the same biological samples were used) and thus follow a 45° line if plotted against each other (Figure 3.2). FDR significant genes on both arrays (blue points) are clearly further from zero and tend to follow the 45° line, with more of those genes having a positive log fold change (and hence up-regulated in Col-0) on both arrays. Points in the upper left and lower right quadrant are genes that differ in the sign of their log fold change between arrays. For example, genes in the lower right quadrant have a positive log fold change in the tiling array, but a negative log fold change in the ATH1 array. Investigating the significant points in this quadrant can help explain why many more down-regulated genes are identified in the ATH1 analysis than in the tiling analysis.

Two other quantities in addition to the mean log fold change affect the significance of a gene. The number of probes (p) and the mean squared error (MSE) per probe

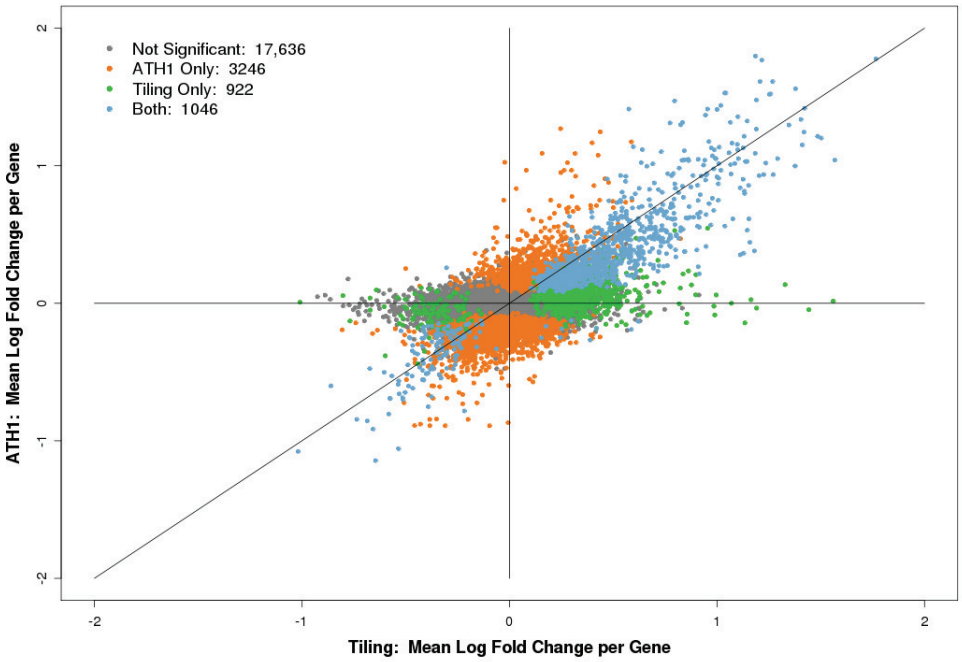

Figure 3.2. Mean log fold change for genes represented on both the ATH1 and tiling arrays. FDR results are in grey (non-significant), orange (significant with ATH1 only), green (significant with tiling only), and blue (significant in both) points. The numbers in the legend correspond to the number of genes in each of the four groups. The 45° line is for comparison purposes.

set also affect the test statistic (3.3) for differential expression. The tiling array has an average of 44 probes per gene, giving it an advantage over the ATH1 array which has an average of 11 probes per probe set. The tiling array, however, also has a larger average MSE (0.884) per probe set than the ATH1 array (0.256). This reduction in variation may be due to the optimal probe selection process in ATH1 arrays.

In summary, several genes are identified as differentially expressed using both arrays and may be of interest for further study. However, even though the same biological samples are hybridized to both array types and the same statistical analysis is used to test for differential expression, there are many discrepancies in the results. Although differences are expected due to the design differences in the two arrays,

this work suggests that more research is needed to better understand the use of tiling arrays for studying differential expression.

3.4 Summary

A common goal of gene expression studies is to identify genes that are differentially expressed between conditions. Gene expression arrays have been widely used for this application for many years. Coupling tiling arrays with their annotation information, an equivalent analysis can be performed. Tiling arrays have a few advantages in this type of analysis, as they offer more coverage per gene and cover more annotated genes. However, because their probes are not specifically designed for the study of gene expression, a comparison between tiling and gene expression arrays for this application is not well understood. This is demonstrated in a real data analysis of *Arabidopsis thaliana* where the same biological samples are hybridized to both array types and the same statistical analysis performed. While the statistical model presented here is relatively simple, it demonstrates how genomic annotation information (i.e., knowing which probes are in exons of genes) can be used to identify biologically relevant probes for a differential expression analysis with tiling arrays.

4. USING GENOMIC ANNOTATION FOR DNA METHYLATION PROFILING WITH TILING ARRAYS

Identifying locations of DNA methylation across a genome is a key step to better understanding the role of this epigenetic mechanism. Tiling arrays allow for the possibility of large-scale (often genome-wide) studies of DNA methylation. These studies present many statistical challenges due to issues such as the large number of probe-level tests (often millions), small number of replicates, dependency between neighboring probes, and experimental noise present in the data. In a typical DNA methylation profiling study, statistical methods are employed to determine which probes or regions of probes are methylated. These results are then visually connected back to genome annotation to identify patterns of DNA methylation for different genomic elements. While investigating the distribution of DNA methylation across the genome is helpful, it may be more meaningful to incorporate genomic annotation into statistical methodology rather than using it after the analysis is complete. In this chapter, current statistical methods for identifying locations of DNA methylation are reviewed and a new approach which extends ideas from these methods to incorporate genomic annotation information into a statistical analysis is proposed.

4.1 Current Statistical Methods

4.1.1 Review of Experimental Procedures and Independent Testing

To understand statistical methods for estimating the DNA methylation status of tiling array probes, it is beneficial to review the experimental procedure and resulting data structure from such studies. Recall that DNA methylation profiling experiments involve the application of treatments such as bisulfite conversion, methylation sen-

sitive restriction enzymes, or methylcytosine immunoprecipitation to genomic DNA samples (Weber et al., 2005; Keshet et al., 2006; Schumacher et al., 2006; Beck and Rakyan, 2008; Estecio and Issa, 2009). As a brief example, consider an experiment which uses the methylation restriction enzyme McrBC to separate methylated and unmethylated DNA (Figure 1.6). A genomic DNA sample from one individual is split into two equal subsamples and digestion with McrBC is applied to one of the subsamples, resulting in a methylation depleted (treated) sample. No treatment is applied to the other subsample and thus it is representative of the total (untreated) genomic DNA with methylation retained. These samples are then hybridized to microarrays. Statistical methods are needed to compare hybridization intensites between the untreated and treated samples for each probe to estimate whether the probe is methylated or not. Methylated probes are expected to have higher hybridization intensities in the untreated sample than the treated sample, since the untreated sample retains methylated DNA and the treated sample does not.

For spotted cDNA arrays, a linear model approach has been successfully applied with an analysis of variance (ANOVA) model (Lippman et al., 2004; Martienssen et al., 2005; Vaughn et al., 2007; Yoo and Doerge, 2009) to test for methylation status of each probe. Recall that for spotted cDNA arrays, both the untreated and treated samples are hybridized to the same array with different dye labels. As with gene expression arrays, probes are usually pre-processed by performing a background correction, normalizing for dye effects, and applying a log-transformation prior to implementation of the statistical model. The ANOVA model (4.1) that has been applied in previous studies and assumes a common variance for all probes is:

$$y_{ijkpr} = \mu + T_i + D_j + A_k + P_p + TP_{ip} + DP_{jp} + AP_{kp} + \epsilon_{ijkpr} \tag{4.1}$$

where $i = 1, 2;\ j = 1, 2;\ k = 1, 2, .., n_a;\ p = 1, ..., n_{up};\ r = 1, ..., n_p;$ and y_{ijkpr} represents the background corrected, normalized, log-transformed intensity of sample type T_i labeled with dye D_j on array A_k for probe P_p and probe replicate r. Note that n_a is the number of arrays, n_{up} is the number of unique probes, and n_p is the number of replicates of probe p. The μ term is the grand mean and T, D, A, P represent the

treatment (sample type), dye, array, and probe main effects, with TP, DP, and AP corresponding to respective interaction effects. The errors, ϵ_{ijkpr}, are assumed to be independent normally distributed random variables with mean 0 and variance σ^2.

The hypotheses (4.2) employed in initial studies to test for methylated probes (Lippman et al., 2004; Martienssen et al., 2005) are:

$$H_o : (T_1 + TP_{1p}) - (T_2 + TP_{2p}) = 0 \quad \text{vs.} \quad H_a : (T_1 + TP_{1p}) - (T_2 + TP_{2p}) \neq 0 \quad (4.2)$$

where T_1 corresponds to the untreated sample and T_2 represents the treated sample. Although these hypotheses (4.2) were proposed and are commonly used for differential expression analysis (Black and Doerge, 2002), DNA methylation profiling studies pose some new issues. First, it is important to note that while conceptually the difference between untreated and treated samples should be near 0 for unmethylated probes (since both samples retain unmethylated DNA), the true effect of the treatment is unknown. To account for any unexpected effect of the treatment other than removing truly methylated DNA, a set of control probes can be chosen which are known *a priori* to be unmethylated. The methylation status of probes can then be determined relative to the control probes. Also, in differential expression analysis, a two-sided test is conducted to indicate that probes can either remain unchanged or be up- or down-regulated. However, for DNA methylation profiling experiments, it is of interest to determine if DNA methylation is either present or absent (relative to control probes) at a particular probe. Thus, a one-tailed test is more appropriate for DNA methylation tiling array data (Yoo and Doerge, 2009).

A set of updated hypotheses (4.3) employed in Vaughn et al. (2007) and detailed in Yoo and Doerge (2009) that address these issues specific to DNA methylation are:

$$H_o : (T_1 + TP_{1p}) - (T_2 + TP_{2p}) \leq \mu_0 \quad \text{vs.} \quad H_a : (T_1 + TP_{1p}) - (T_2 + TP_{2p}) > \mu_0 \quad (4.3)$$

where μ_0=median$\{(T_1 + TP_{1c}) - (T_2 + TP_{2c}), \forall c$ control probes$\}$. The test statistic (4.4) for testing hypotheses (4.3) at probe p is:

$$z_p^* = \frac{(T_1 + TP_{1p}) - (T_2 + TP_{2p}) - \mu_0}{\sqrt{\frac{2\hat{\sigma}^2}{n_a n_p}}} \sim N(0,1) \quad \text{under } H_0 \qquad (4.4)$$

Table 4.1
Example of data from a DNA methylation Affymetrix® tiling array experiment. y_{ipk} represents the background corrected, normalized, log-transformed intensity of the p^{th} probe, for sample type (untreated or treated) i, of biological replicate k, where $i = 1, 2$; $p = 1, ..., P; k = 1, 2, .., n$. d_{pk} is the paired difference between the untreated and treated sample collected from the k^{th} individual at probe p.

Biological Replicate	Untreated (Total DNA)	Treated (Methyl Depleted)	Paired Difference
1	y_{1p1}	y_{2p1}	$d_{p1} = y_{1p1} - y_{2p1}$
2	y_{1p2}	y_{2p2}	$d_{p2} = y_{1p2} - y_{2p2}$
3	y_{1p3}	y_{2p3}	$d_{p3} = y_{1p3} - y_{2p3}$
...	...	...	...
n	y_{1pn}	y_{2pn}	$d_{pn} = y_{1pn} - y_{2pn}$

where $\hat{\sigma}^2$ is the common variance estimate for all probes. Note that a per-probe variance could also be employed, with appropriate adjustments to the test statistic and its distribution under the null hypothesis.

Affymetrix® DNA methylation tiling array experiments differ from spotted cDNA arrays in that each of the two samples (untreated and treated) collected from the same individual is hybridized to a separate array and this process is repeated for additional biological replicates. Affymetrix® tiling array data are typically pre-processed by background correction, normalization to remove array effects, and log-transformation prior to employing statistical methods for sample comparison. Table 4.1 gives an example of the pre-processed data generated from such an experiment using Affymetrix® tiling arrays.

With data in this format, a simplistic approach for determining whether a probe is significantly methylated is to conduct a paired t-test at each probe. Pairwise differences (Table 4.1) can be calculated between observations on the same individual

under the two different conditions, $d_{pk} = y_{1pk} - y_{2pk}$. This difference is expected to be large for methylated probes since the untreated sample retains methylated DNA and the treated sample does not. The following hypotheses (4.5), which address the issues specific to DNA methylation tiling array experiments, can be tested for each probe:

$$H_o : \mu_{d_p} \leq \mu_0 \qquad \text{vs.} \qquad H_a : \mu_{d_p} > \mu_0 \tag{4.5}$$

where μ_{d_p} is the population mean for the paired differences between the untreated and treated samples for probe p and μ_0=median$\{\mu_{d_c}, \forall c$ control probes$\}$. The test statistic (4.6) for each probe is:

$$t_p^* = \frac{\overline{d}_p - \mu_0}{\frac{S_{d_p}}{\sqrt{n}}} \sim t(n-1) \qquad \text{under} \;\; H_o \tag{4.6}$$

where $\overline{d}_p = \frac{\sum_{k=1}^{n} d_{pk}}{n}$ and $S_{d_p} = \sqrt{\frac{\sum_{k=1}^{n} (d_{pk} - \overline{d}_p)^2}{n-1}}$, noting that S_{d_p} may differ for each probe (i.e., per-probe variance). Note that a common probe variance could also be employed with appropriate adjustments to the test statistic and its distribution under the null hypothesis.

For both spotted cDNA and Affymetrix® tiling arrays, the multiple testing issue that arises due to the large number of hypothesis tests being conducted is typically addressed by controlling the false discovery rate (FDR) at level α (Benjamini and Hochberg, 1995). This procedure for controlling the FDR assumes the hypotheses tests conducted for each probe are independent. However, probes are linearly ordered across a genomic region, making 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), offering further evidence that the methylation status of probes may depend on neighboring probes. Effectively utilizing information from neighboring probes is a current statistical challenge for DNA methylation profiling studies with tiling arrays and are the main focus of this research.

4.1.2 Methods Incorporating Dependency Between Probes

An initial effort to incorporate the potential dependency between neighboring tiling array probes into a statistical model involved the implementation of sliding window tests to identify transcription factor binding sites (Cawley et al., 2004; Ji and Wong, 2005; Buck et al., 2005; Keles et al., 2006). These methods combine information from probes within a certain genomic distance of the probe being tested to calculate a test statistic for that probe. Test statistics employed vary for different procedures. For example, Cawley et al. (2004) use all probes within a window of 1000 bases of the probe being tested to calculate a Wilcoxon rank sum statistic. Keles et al. (2006) allow a varying genomic window size so that the same number of probes will be included in each window and probes in the window are combined via a moving average (MA) of t-test statistics. While these methods accommodate the spatial structure of the probes, they also present a few problems. First, if the window size is based on genomic distance, test statistics for some probes may only utilize information from a small number of probes; whereas if the window size is based on a fixed number of probes then the genomic distance between probes in the window may be large so that a dependency between some of the probes in the window is unlikely. Also, a separate hypothesis test is still conducted for each probe, and although Keles et al. (2006) proposed some options for accommodating this multiple testing problem when tests are not independent, these alternatives are not optimal and testing of multiple dependent probes remains an open problem for these methods.

An alternative framework to conducting individual hypotheses tests for each probe, which also incorporates dependency between neighboring tiling array probes is a hidden Markov model (HMM) (Rabiner, 1989; Cappe et al., 2005). Hidden Markov models have been proposed to identify locations of transcription factor binding sites and histone modifications through ChIP-chip experiments (Li et al., 2005; Ji and Wong, 2005; Du et al., 2006; Humburg et al., 2008). Yoo (2008) also proposed an HMM for detecting locations of DNA methylation. In a hidden Markov model, a sequence of

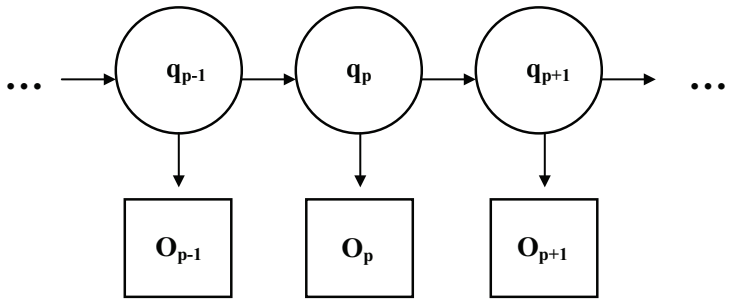

Figure 4.1. Illustration of a hidden Markov model. The random variable q_p represents the hidden state at probe p, while the random variable O_p represents the observed value at probe p. Arrows represent conditional dependencies. Since the hidden state (q_p) at probe p only depends on the hidden state (q_{p-1}) at probe $p-1$, the first-order Markov property holds. Also, the observation (O_p) at probe p is conditionally dependent on the hidden state (q_p) at probe p .

non-observable (hidden) random variables take on values in a set of finite states and form a Markov chain. Although the states themselves are not directly observable, an observable output is available, which is dependent on the hidden states (Figure 4.1). In the case of DNA methylation tiling array experiments, the hidden states are the true methylation status of the probes (methylated or not) and the observed values are the intensity measurements. Standard algorithms (forward-backward (Baum et al., 1970; Baum, 1972), Viterbi (Viterbi, 1967; Forney, 1973), Baum-Welch (Baum et al., 1970)) are available for HMMs which can estimate the model parameters and the hidden states using information from all probes.

A popular two-stage HMM approach for modeling tiling array data was proposed by Ji and Wong (2005), where the first stage involves the calculation of empirical Bayes t-statistics for each probe. In the second stage, the hidden states for each probe are estimated with the forward-backward algorithm (Baum et al., 1970; Baum, 1972) using probe-level statistics as observed values. The algorithm in the second stage is restarted if two neighboring probes are more than a specified genomic distance apart. Ji and Wong (2005) developed the TileMap software to implement this method and

it has been used in a variety of studies, including the identification of locations of DNA methylation (Zhang et al., 2006) and histone modifications (Zhang et al., 2007) in *Arabidopis thaliana*.

Although the approach presented by Ji and Wong (2005) provides an advantageous way to incorporate neighboring probes into DNA methylation status estimation, Humburg et al. (2008) pointed out that some of the model parameters in Ji and Wong (2005) and other methods are estimated in an *ad hoc* manner. This is also the case for the HMM proposed by Li et al. (2005), who use results from a previous study, and Yoo (2008), who uses the observed methylation status obtained from a hypothesis test to estimate HMM model parameters. Munch et al. (2006) and Du et al. (2006) use genomic annotation to identify a set of training data, which can be used to estimate the HMM model parameters. Both studies focus on transcription mapping, where expressed and non-expressed regions are identified, although Du et al. (2006) notes a potential application to transcription factor binding site identification and posits that any source of validated biological knowledge can be used in the context of the particular application to select a good set of training data. Humburg et al. (2008) propose the use of the Baum-Welch (Baum et al., 1970) or Viterbi training (Juang and Rabiner, 1990) algorithms to obtain maximum likelihood estimates for HMM model parameters using data from all probes rather using *ad hoc* methods. They apply this method to a ChIP-chip histone modification tiling array study. As an alternative to HMMs, a few Bayesian approaches have also been proposed for modeling tiling array data (Qi et al., 2006; Keles, 2007; Gottardo et al., 2008; Wu et al., 2009; Mo and Liang, 2010). Although these methods are potentially powerful alternatives, some of them require experimental information that are not always readily available or are computationally more complex than the standard HMM algorithms.

While the previous methods highlight the importance of incorporating a dependency structure among probes and using a formal estimation procedure for model parameters, the majority of them do not make use of genomic annotation information. The only exceptions are the methods by Du et al. (2006) and Munch et al.

(2006) which utilize genomic annotation to identify a set of probes to be used for training data. Unfortunately, both of these methods used the training data to arrive at a set of common parameter estimates for the whole genomic region, and neither were applied to DNA methylation tiling array data. DNA methylation status of tiling array probes is typically identified using one of the statistical methods described in this section and then results are connected back to genome annotation to identify patterns after the analysis.

Genomic annotation for many organisms is now available and previous studies have revealed that, for some organisms, different genomic elements (e.g., genes, transposons) may have different DNA methylation patterns than other genomic regions (Suzuki and Bird, 2008). This research is motivated by a desire to improve DNA methylation status prediction by coupling knowledge about genomic annotation with statistical analysis for DNA methylation tiling array data. Here, a method is proposed to incorporate genomic annotation information into a HMM framework for estimating the DNA methylation status of tiling array probes. In particular, certain model parameters are allowed to vary according to the genomic element the probe represents and HMM estimation procedures are modified to include this extra layer of information. The resulting model integrates the use of neighboring probe dependency with genomic annotation, while obtaining maximum likelihood estimates of HMM model parameters.

4.2 Hidden Markov Models (HMMs)

As described previously, a hidden Markov model is a probabilistic process in which a set of underlying, unobservable hidden states form a Markov chain (Cappe et al., 2005). At each time point (probe) in the chain, an observation is available whose distribution depends on the underlying state (Figure 4.1). Prior to their successful application in epigenomic studies, HMMs have been employed in a variety of applications from speech recognition (Rabiner, 1989; Jelinek, 1997) to analysis of biological

sequence data (Durbin et al., 1998). In this section, the general HMM framework is introduced along with a description of standard estimation algorithms and is discussed in the context of DNA methylation profiling studies with tiling arrays.

4.2.1 General HMM Framework

The following summary about HMMs is based upon the model formulation in a tutorial by Rabiner (1989). For hidden states $S = \{S_1, ..., S_N\}$ a sequence of random variables $\{q_p, p = 1, ..., P\}$ take on values in S. The variable q_p represents the state the process is in at time (or probe) p. The first-order Markov property holds for this sequence such that the probability that the process is in state S_j at probe $p+1$ (i.e., $q_{p+1} = S_j$) depends only on the state at probe p and no other previous states:

$$\begin{aligned} a_{ij} &= P(q_{p+1} = S_j | q_p = S_i, q_{p-1} = S_{i_{p-1}}, ..., q_1 = S_{i_1}) \\ &= P(q_{p+1} = S_j | q_p = S_i). \end{aligned} \tag{4.7}$$

These probabilities (4.7) are transition probabilities which provide the probability of the process going from state S_i at probe p to state S_j at probe $p+1$ for any p (i.e., independent of time/probe). Transition probabilities (a_{ij}) have the following properties (4.8):

$$\begin{aligned} a_{ij} &\geq 0 \\ \Sigma_{j=1}^{N} a_{ij} &= 1 \qquad\qquad 1 \leq i, j \leq N \end{aligned} \tag{4.8}$$

and can be written in matrix form $A = \{a_{ij}\}$:

$$A = \begin{pmatrix} a_{11} & a_{12} & ... & a_{1N} \\ a_{21} & a_{22} & ... & a_{2N} \\ ... & ... & ... & ... \\ a_{N1} & a_{N2} & ... & a_{NN} \end{pmatrix}. \tag{4.9}$$

The probability of the first probe being in state S_i is given by the initial state distribution $\pi = \{\pi_i\}$:

$$\pi_i = P(q_1 = S_i) \qquad 1 \leq i \leq N. \tag{4.10}$$

Although the hidden states are unobservable, at each probe p an observation (O_p) is available (which may be discrete or continuous) that depends on the hidden state at probe p. This is given by the observation probability distribution in state j, $B = \{b_j(o_p)\}$:

$$b_j(o_p) = f(o_p|q_p = S_j) \qquad 1 \leq j \leq N, \quad -\infty < o_p < \infty. \tag{4.11}$$

A hidden Markov model is characterized by hidden states (S), the state transition probability distribution (A), the initial probability distribution (π), and the observation probability distribution (B). The complete parameter set (4.12) of the model is denoted:

$$\lambda = (A, B, \pi). \tag{4.12}$$

To simulate a sequence of P hidden states $Q = \{q_1, q_2, ..., q_P\}$ and observations $O = \{o_1, o_2, ..., o_P\}$ from an HMM, the following steps can be implemented, given parameter values for λ:

1. Use the initial probability distribution (4.10) to select the hidden state at the first probe ($q_1 = S_i$).

2. Let $p = 1$.

3. Use the observation probability distribution (4.11) in state S_i to obtain an observed value o_p.

4. Use the state transition probability distribution (4.9) for state S_i to determine the next hidden state $q_{p+1} = S_j$.

5. For $p < P$, set $p = p + 1$ and go back to step 3. Once the final hidden state q_P and observation o_P are generated, the procedure is stopped.

4.2.2 HMMs for DNA Methylation Profiling

For DNA methylation tiling array experiments, the goal is to determine whether each of the $p = \{1, ..., P\}$ probes is methylated or not. However, the true methylation

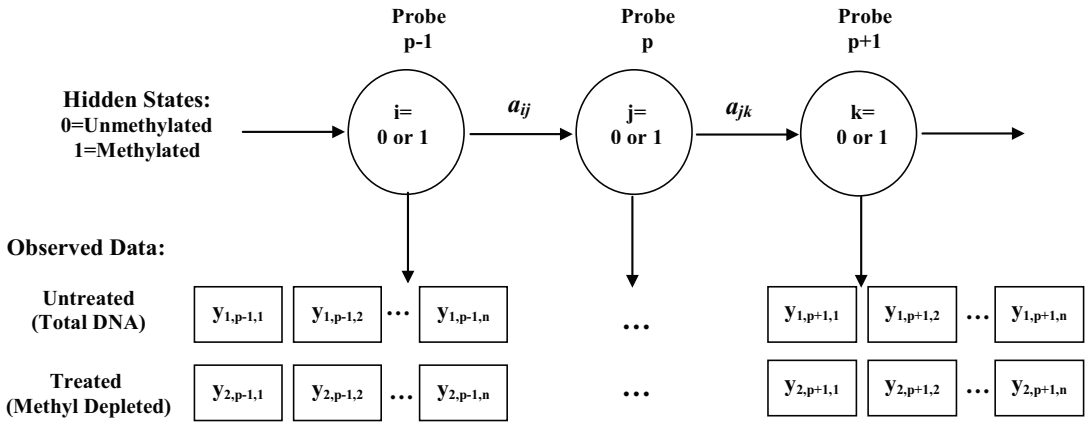

Figure 4.2. A hidden Markov model for DNA methylation profiling using tiling arrays. The circles represent probes with hidden states 0 for unmethylated probes and 1 for methylated probes. Arrows represent conditional dependencies. The hidden states for the probes follow a first-order Markov chain with 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. The boxes represent the observations (y_{ipk}) which are background corrected, normalized, log-transformed intensities from the tiling array experiment, where $i = \{1, 2\}$ is the sample type (untreated, treated), $p = \{1, ..., P\}$ is the probe and $k = \{1, ..., n\}$ is the biological replicate.

status is unknown and, instead, the information actually observed for each probe are intensity values as in Table 4.1 for Affymetrix® arrays. For spotted cDNA arrays, the experimental effects (dye, array) can be removed to obtain data in the same format. Thus, in terms of the HMM framework, the hidden states are $S = \{0, 1\}$ where 0 is the unmethylated state and 1 is the methylated state. Figure 4.2 illustrates the hidden Markov model in the context of DNA methylation profiling. The initial probabilities $\pi = \{\pi_0, \pi_1\}$ can be thought of as the proportion of probes that are unmethylated or methylated in the genomic region of interest. The transition matrix (4.13):

$$A = \begin{pmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{pmatrix} \tag{4.13}$$

gives the probabilities of moving from methylation status {0,1} at probe p to methylation status {0,1} at probe $p+1$ (e.g., a_{00} gives the probability of a probe being unmethylated at probe p and staying unmethylated at probe $p+1$). For both initial probabilities and transition probabilities, information can be used from previous studies, or empirically estimated from the data, to obtain reasonable starting parameter estimates for a formal estimation procedure. For instance, it is expected that a state is more likely to remain the same upon transition than to change (i.e., a_{00} and a_{11} are expected to be much greater than a_{01} and a_{10}).

Finally, the observed data for each probe consist of intensity values of untreated and treated samples for n biological replicates. Since the data are paired (i.e., untreated and treated samples come from the same individual), it is reasonable to assume the paired data (y_{1pk}, y_{2pk}) for each probe p and individual k follow a bivariate normal distribution, conditional on the true methylation status. In other words, the observation probability distribution (4.14) for truly unmethylated (b_0) and methylated (b_1) probes is assumed to be:

$$\begin{aligned} b_0 \begin{pmatrix} y_{1pk} \\ y_{2pk} \end{pmatrix} &\sim N\left(\begin{pmatrix} \mu_{01} \\ \mu_{02} \end{pmatrix}, \begin{pmatrix} \sigma_{01}^2 & \rho_0\sigma_{01}\sigma_{02} \\ \rho_0\sigma_{01}\sigma_{02} & \sigma_{02}^2 \end{pmatrix} \right) \quad \text{Unmethylated Probes} \\ b_1 \begin{pmatrix} y_{1pk} \\ y_{2pk} \end{pmatrix} &\sim N\left(\begin{pmatrix} \mu_{11} \\ \mu_{12} \end{pmatrix}, \begin{pmatrix} \sigma_{11}^2 & \rho_1\sigma_{11}\sigma_{12} \\ \rho_1\sigma_{11}\sigma_{12} & \sigma_{12}^2 \end{pmatrix} \right) \quad \text{Methylated Probes.} \end{aligned} \tag{4.14}$$

The means for both the untreated and treated sample are expected to be similar in magnitude for probes that are not methylated (i.e., $\mu_{01} \approx \mu_{02}$). For methylated probes, the mean of the untreated sample is expected to be larger than the mean of the treated sample (i.e., $\mu_{11} > \mu_{12}$). Since the untreated and treated samples represent DNA collected from the same individual, correlation (ρ_0 and ρ_1) between these observations is expected. Since the parameters for these distributions are unknown in practice, they can be empirically estimated from the real data to obtain starting estimates for a formal estimation procedure.

An additional consideration is that the general HMM framework includes only one observation per probe. However, in this context, there are n observations per probe. To stay within the general HMM framework, one option is to average the n observations, leading to the following modified observation probability distribution (4.15):

$$\begin{aligned} b_0\begin{pmatrix}\overline{y}_{1p.}\\ \overline{y}_{2p.}\end{pmatrix} &\sim N\left(\begin{pmatrix}\mu_{01}\\ \mu_{02}\end{pmatrix}, \frac{1}{n}\begin{pmatrix}\sigma_{01}^2 & \rho_0\sigma_{01}\sigma_{02}\\ \rho_0\sigma_{01}\sigma_{02} & \sigma_{02}^2\end{pmatrix}\right) \quad \text{Unmethylated Probes} \\ b_1\begin{pmatrix}\overline{y}_{1p.}\\ \overline{y}_{2p.}\end{pmatrix} &\sim N\left(\begin{pmatrix}\mu_{11}\\ \mu_{12}\end{pmatrix}, \frac{1}{n}\begin{pmatrix}\sigma_{11}^2 & \rho_1\sigma_{11}\sigma_{12}\\ \rho_1\sigma_{11}\sigma_{12} & \sigma_{12}^2\end{pmatrix}\right) \quad \text{Methylated Probes.} \end{aligned} \tag{4.15}$$

Another option is to combine information from multiple observation sequences into the state and parameter estimation algorithms. This alternative is discussed in the future work chapter, but for the remainder of this research the average of the n observations is employed.

4.2.3 HMM Estimation Algorithms

For hidden Markov models, there are three main problems that need to be addressed. First, the likelihood function, or the probability of the observation sequence given the model parameters $P(O|\lambda)$, needs to be efficiently computed. Second, the hidden states need to be estimated in an optimal way given the observation sequence and the model parameters. And, finally, the model parameters need to be estimated so that $P(O|\lambda)$ is maximized. Standard algorithms have been developed to address each of these issues (Rabiner, 1989). The forward-backward algorithm (Baum et al., 1970; Baum, 1972) can both efficiently compute $P(O|\lambda)$ and estimate the hidden states. When used to estimate the hidden states, the forward-backward algorithm maximizes the expected number of correct individual states. An alternative algorithm for estimating hidden states is the Viterbi algorithm (Viterbi, 1967; Forney,

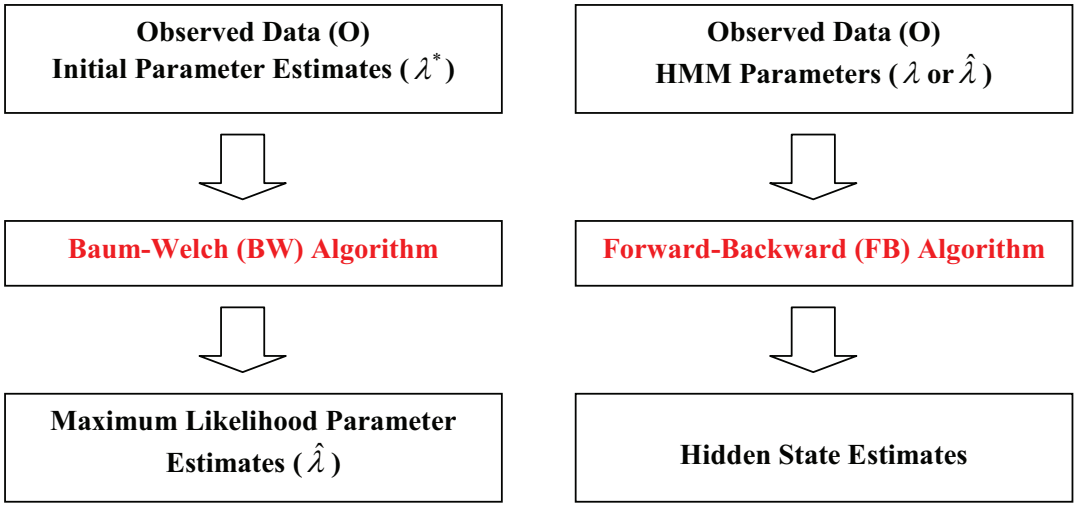

Figure 4.3. Workflow of the Baum-Welch and forward-backward algorithms. The Baum-Welch algorithm requires a set of observations (O) and initial parameter estimates (λ^*) as inputs to calculate the maximum likelihood parameter estimates ($\hat{\lambda}$). The forward-backward algorithm requires a set of observations (O) and model parameters, which can be the true parameters (λ) if they are known or the maximum likelihood parameter estimates ($\hat{\lambda}$) obtained from the Baum-Welch algorithm. With this information, the forward-backward algorithm estimates the hidden states.

1973), which maximizes the probability of the state sequence (rather than individual states). While both algorithms and optimality criteria have their own merits, in this work, the percent of individual correctly predicted states will be used as a model performance measure and, thus, the forward-backward algorithm will be employed. Maximum likelihood estimates for the model parameters can be obtained through the Baum-Welch algorithm (Baum et al., 1970; Baum, 1972), which is the expectation-maximization (EM) algorithm (Dempster et al., 1977) for HMMs. Details of both the forward-backward and the Baum-Welch algorithm are described in this section, since they are both used in this work. The forward-backward procedure is described first since it requires the creation of forward and backward variables which are also used in the Baum-Welch algorithm. Figure 4.3 gives an overview of these methods.

Forward-Backward (FB) Algorithm

One of the main goals when using a HMM is to obtain an optimal estimate of the hidden state sequence (Q) given the observations (O) and the model parameters (λ). In the context of DNA methylation tiling array experiments, this translates to estimating the DNA methylation status of all probes given the tiling array intensity data for each probe and a specific set of model parameters. The FB algorithm (Baum et al., 1970; Baum, 1972) accomplishes this goal by utilizing the probability (4.16) of being in state S_i at probe p, given the observation sequence O and the model parameters λ:

$$\gamma_p(i) = P(q_p = S_i|O, \lambda). \tag{4.16}$$

For each probe p, the state which yields the maximum probability $\gamma_p(i)$ is the estimated hidden state for that probe. Thus, the FB algorithm selects the states that are individually most likely, maximizing the expected number of correct individual states (Rabiner, 1989).

The probability $\gamma_p(i)$ is calculated through the use of forward and backward variables. The forward variable (4.17) is the joint probability of the partial observation sequence up to probe p and state S_i at probe p given the model λ:

$$\alpha_p(i) = P(o_1 o_2 ... o_p, q_p = S_i|\lambda). \tag{4.17}$$

The forward variables $\alpha_p(i)$ can be calculated inductively:

1. Initial Step:
   $\alpha_1(i) = \pi_i b_i(o_1) \qquad 1 \leq i \leq N$

2. Induction Step:
   $\alpha_{p+1}(j) = [\sum_{i=1}^{N} \alpha_p(i) a_{ij}] b_j(o_{p+1}) \quad 1 \leq j \leq N, 1 \leq p \leq P-1.$

Figure 4.4 illustrates the inductive step of the forward variable calculation for a HMM with two states {0,1}, as is the case for DNA methylation tiling array data. Note

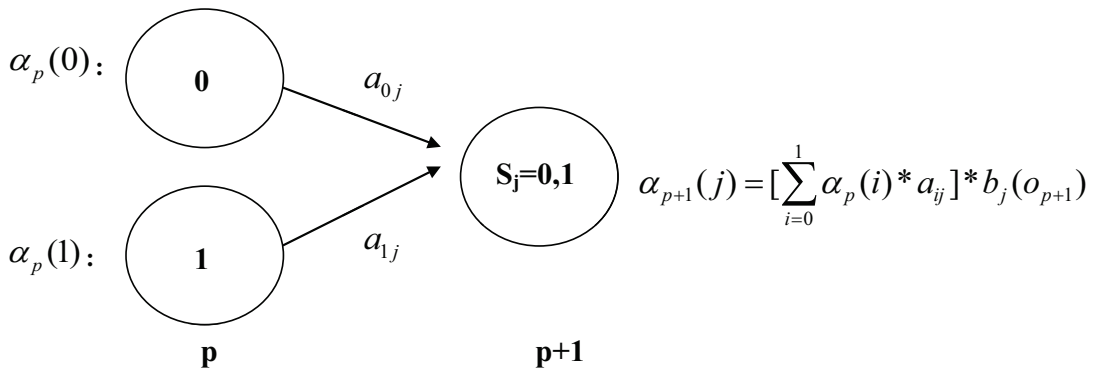

Figure 4.4. The inductive step in the forward variable ($\alpha_{p+1}(j)$) calculation for a HMM with two states $S_i = \{0, 1\}$. Probe $p + 1$ could have arrived at state S_j either through the hidden state 0 or 1 at probe p. The forward variables for probe p ($\alpha_p(0)$ and $\alpha_p(1)$) represent the joint probability of the partial observation sequence up to probe p and state S_i at probe p. Thus the product $\alpha_p(i) * a_{ij}$ is the joint probability of the partial observation sequence up to probe p and reaching state S_j at probe $p+1$ through state S_i at probe p. Summing across these probabilities and accounting for the observation at probe $p + 1$ by multiplying the sum by $b_j(o_{p+1})$ gives the joint probability of the partial sequence up to probe $p + 1$ and state S_j at probe $p + 1$ (Rabiner, 1989).

that $P(O|\lambda)$ can be calculated (4.18) by summing over the forward variables for the last probe (P) in the sequence (Rabiner, 1989):

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_P(i). \tag{4.18}$$

The backward variable (4.19) represents the probability of the partial observation sequence from probe $p + 1$ to the end of the sequence, given state S_i at probe p and the model parameters λ:

$$\beta_p(i) = P(o_{p+1} o_{p+2} ... o_P | q_p = S_i, \lambda). \tag{4.19}$$

Similar to the forward variables, the backward variables $\beta_p(i)$ can also be calculated inductively (Rabiner, 1989):

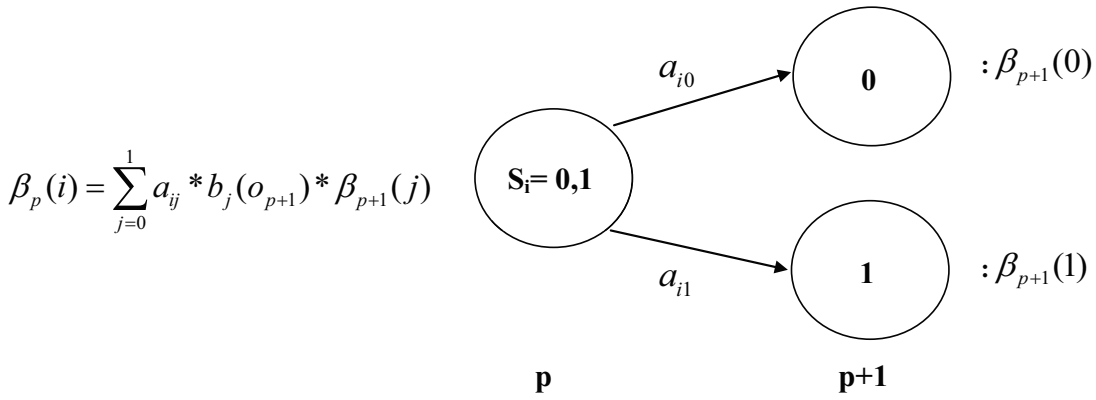

Figure 4.5. The inductive step in the backward variable ($\beta_p(i)$) calculation for a HMM with two states $S_i = \{0, 1\}$. If the hidden state at probe p is S_i, then a transition to either state 0 or 1 at probe $p+1$ can occur. The backward variables at probe $p+1$ ($\beta_{p+1}(0)$ and $\beta_{p+1}(1)$) represent the probability of the partial observation sequence from probe $p+2$ to the end of the sequence, given state S_j at probe $p+1$. Thus the product $a_{ij} * b_j(o_{p+1}) * \beta_{p+1}(j)$ accounts for the transition (a_{ij}) from state S_i at probe p to state S_j at probe $p+1$, the observation at $p+1$ ($b_j(o_{p+1})$), and the observations from $p+2$ to the end of the sequence ($\beta_{p+1}(j)$). Summing this product over all possible states S_j gives the probability of the partial observation sequence from probe $p+1$ to the end of the sequence, given state S_i at probe p (Rabiner, 1989).

1. Initialization Step:

   $\beta_P(i) = 1 \qquad 1 \leq i \leq N$

2. Induction Step:

   $\beta_p(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{p+1}) \beta_{p+1}(j) \qquad 1 \leq i \leq N, p = P-1, P-2, ...1.$

Figure 4.5 highlights the inductive step calculation of the backward variable for a HMM with two states {0,1}, as is the case for DNA methylation tiling array data.

Finally, the forward and backward variables can be employed to calculate $\gamma_p(i)$ as follows (4.20):

$$\gamma_p(i) = P(q_p = S_i | O, \lambda) = \frac{\alpha_p(i)\beta_p(i)}{\sum_{i=1}^{N} \alpha_p(i)\beta_p(i)}. \tag{4.20}$$

Thus, the individually most likely state at probe p (4.21) (Rabiner, 1989):

$$q_p = \arg\max_{1 \leq i \leq N}[\gamma_p(i)] \qquad 1 \leq p \leq P. \tag{4.21}$$

Note that $\gamma_p(i)$ gives an certainty measure for the state estimate at probe p. For example, if state S_i maximizes $\gamma_p(i)$, then more confidence can be placed in the state estimate when $\gamma_p(i)$ is closer to 1 than to 0.5. This extra information obtained from the forward-backward algorithm, along with the fact that it maximizes the expected number of correctly predicted states, make it the state estimation algorithm of choice for this work.

Baum-Welch (BW) Algorithm

While the forward-backward algorithm can provide DNA methylation status estimates for all probes, it requires the knowledge of a set of model parameters λ. In real data, the true model parameters are unknown and need to be estimated from the data. Ideally, the model parameters λ should maximize the probability of the observation sequence given the model parameters ($P(O|\lambda)$). While there is no analytical solution to find this maximum, the expectation-maximization (EM) algorithm (Dempster et al., 1977) can be used to find λ such that $P(O|\lambda)$ is locally maximized. Baum et al. (1970) formalized this procedure within the HMM framework and it is called the Baum-Welch (BW) algorithm. Given a set of initial parameters and the observed data, the BW algorithm can re-estimate the parameters of the initial state (π), transition (A), and observation probability (B) distributions until a local maximum of $P(O|\lambda)$ is reached.

The BW algorithm makes use of the forward (4.17) and backward (4.19) variables, as well as $\gamma_p(i)$ (4.16), described in the formulation of the forward-backward

algorithm. Additionally, the following probability (4.22) is also employed in the BW algorithm:

$$\begin{aligned}\xi_p(i,j) &= P(q_p = S_i, q_{p+1} = S_j|O,\lambda) \\ &= \frac{\alpha_p(i)a_{ij}b_j(o_{p+1})\beta_{p+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_p(i)a_{ij}b_j(o_{p+1})\beta_{p+1}(j)}\end{aligned} \tag{4.22}$$

which represents the probability of being in state S_i at probe p and in state S_j at probe $p+1$, given the observations and model parameters. Given a set of initial parameter estimates $\lambda^* = (A^*, B^*, \pi^*)$ and the observed data, the following re-estimation formulas for the initial and transition probabilities can be applied:

Initial Probabilities:

$$\begin{aligned}\hat{\pi}_i &= \gamma_1(i) \\ &= \text{Expected number of times in state } S_i \text{ at the first probe}\end{aligned} \tag{4.23}$$

Transition Probabilities:

$$\begin{aligned}\hat{a}_{ij} &= \frac{\sum_{p=1}^{P-1}\xi_p(i,j)}{\sum_{p=1}^{P-1}\gamma_p(i)} \\ &= \frac{\text{Expected number of transitions from state } S_i \text{ to state } S_j}{\text{Expected number of transitions out of state } S_i}.\end{aligned} \tag{4.24}$$

Baum et al. (1970) also proposed a re-estimation formula for the observation probability distribution parameters for discrete observations, yielding an updated set of parameter estimates $\hat{\lambda} = (\hat{A}, \hat{B}, \hat{\pi})$ derived from the re-estimation formulas. That work showed that $P(O|\hat{\lambda}) \geq P(O|\lambda^*)$, meaning the observations are more likely to have been produced by the model with updated parameters $\hat{\lambda}$ than the model with the initial parameters λ^*. This re-estimation procedure can be iteratively applied by continually inputing the new set of parameters into the re-estimation formulas until some convergence criteria is met (i.e., $P(O|\hat{\lambda}) - P(O|\lambda^*) < \epsilon, \epsilon > 0$)(Rabiner, 1989).

Since the observations in a DNA methylation profiling experiment are continuous rather than discrete, a different set of parameter estimates than those proposed by Baum et al. (1970) are required for the observation probability distribution. Rabiner (1989) provides details of re-estimation formulas for the parameters of a mixture of normal distributions. For the bivariate normal distributions (4.14, 4.15) that are

assumed for observations from DNA methylation data from tiling arrays, the following re-estimation formulas (4.25) are applicable:

Observation Probability Distribution Parameters:

$$\hat{\mu}_i = \frac{\sum_{p=1}^{P} \gamma_p(i) o_p}{\sum_{p=1}^{P} \gamma_p(i)} \qquad \hat{\Sigma}_i = \frac{\sum_{p=1}^{P} \gamma_p(i)(o_p - \mu_i)(o_p - \mu_i)^T}{\sum_{p=1}^{P} \gamma_p(i)}. \tag{4.25}$$

Taken along with the initial (4.23) and transition (4.24) probability estimates, this re-estimation procedure yields maximum likelihood estimates for all of the HMM parameters. The resulting parameter estimates $\hat{\lambda}$ can then be used in the FB algorithm to obtain hidden state estimates. Note that upon implementation, a scaling procedure is required when the number of probes exceeds 100 to perform calculations within the precision range of a computer.

4.3 Incorporating Genomic Annotation into HMM Framework

In an effort to include additional genomic information into the estimation of DNA methylation status, the hidden Markov model framework is extended to incorporate information provided by genomic annotation. A method is proposed and investigated to determine if incorporating genomic annotation into a HMM framework is useful in predicting true methylation status of tiling array probes in DNA methylation profiling studies. Recall that previous studies (Zhang et al., 2006; Suzuki and Bird, 2008) have found evidence for DNA methylation patterns that differ by genomic element (e.g., gene vs. intergenic regions). For example, DNA methylation studies of *Arabidopsis thaliana* identified regions of dense DNA methylation in bodies of genes interspersed between regions of little or no methylation. In the HMM framework, this can be modeled by allowing probes in genes to have different transition probabilities than probes in intergenic regions (Figure 4.6). To formally incorporate the potential differences in transition probabilities for genes and intergenic regions into the HMM framework, modifications to the forward-backward and Baum-Welch algorithms are necessary. Since both algorithms rely on the use of forward and backward variables, the modifications start at that step.

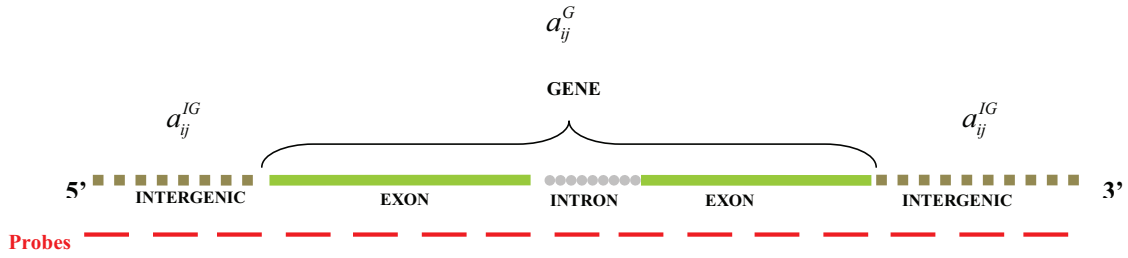

Figure 4.6. Example of how genomic annotation can be incorporated into the HMM framework for DNA methylation tiling array experiments. Probes that correspond to gene regions can have different transition probabilities (a_{ij}^{G}) than probes in intergenic regions (a_{ij}^{IG}) to reflect different dependency patterns in those regions.

4.3.1 Modified Forward and Backward Variables

Recall the forward (4.17) and backward (4.19) variables described previously and the inductive step of their calculations (repeated here for convenience):

1. Forward Variable:
$$\alpha_p(j) = P(o_1 o_2 ... o_p, q_p = S_j|\lambda)$$
$$= [\sum_{i=1}^{N} \alpha_{p-1}(i) a_{ij}] b_j(o_p)$$

2. Backward Variable:
$$\beta_p(i) = P(o_{p+1} o_{p+2} ... o_P | q_p = S_i, \lambda)$$
$$= \sum_{j=1}^{N} a_{ij} b_j(o_{p+1}) \beta_{p+1}(j).$$

To incorporate transition probabilities that vary for genes and intergenic regions, the calculations of the forward and backward probabilities are mostly the same with a_{ij} replaced by a_{ij}^{IG} for intergenic probes and by a_{ij}^{G} for probes in genes. However, due to the first-order Markovian property, the hidden state at probe p depends on the hidden state at probe $p-1$. So, at boundaries where an intergenic region changes to a gene region (or vice versa) the boundary transition probability (a_{ij}^{B}) is the probability

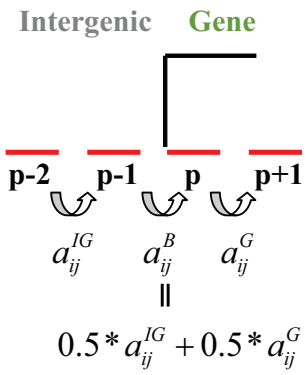

Figure 4.7. Probes at the boundary of an intergenic region and a gene. The transition probability from probe $p-2$ to probe $p-1$ is given by a_{ij}^{IG}, since both probes lie in the intergenic region. Similarly, the transition probability from probe p to probe $p+1$ is given by a_{ij}^{G}, since both probes lie in the gene region. However, the transition probability at the boundary, from probe $p-1$ to probe p, is an average of the intergenic and gene transition probabilities: $a_{ij}^{B} = 0.5 * a_{ij}^{IG} + 0.5 * a_{ij}^{G}$. This transition also occurs when going from a gene to an intergenic region.

of a transition from state S_i at probe $p-1$ in an intergenic region to state S_j at probe p in a gene region (Figure 4.7). This boundary transition probability is taken to be the average of the transition probabilities for gene and intergenic regions (i.e., $a_{ij}^{B} = 0.5 * a_{ij}^{IG} + 0.5 * a_{ij}^{G}$).

To illustrate the changes in the forward and backward variable calculations at boundary regions, consider probes $p-2$ and $p-1$ at the end of an intergenic region and probes p and $p+1$ at the beginning of a gene region, as in Figure 4.7:

Forward Variable Modification:

At probe $p-1$: $\alpha_{p-1}(j) = [\sum_{i=1}^{N} \alpha_{p-2}(i) a_{ij}^{IG}] b_j(o_{p-1})$

At probe p: $\alpha_{p}(j) = [\sum_{i=1}^{N} \alpha_{p-1}(i)(0.5 * a_{ij}^{IG} + 0.5 * a_{ij}^{G})] b_j(o_{p})$

At probe $p+1$: $\alpha_{p+1}(j) = [\sum_{i=1}^{N} \alpha_{p}(i) a_{ij}^{G}] b_j(o_{p+1})$

Backward Variable Modification:

At probe p: $\beta_p(i) = \sum_{j=1}^{N} a_{ij}^{G} b_j(o_{p+1})\beta_{p+1}(j)$

At probe $p-1$: $\beta_{p-1}(i) = \sum_{j=1}^{N} (0.5 * a_{ij}^{IG} + 0.5 * a_{ij}^{G}) b_j(o_p)\beta_p(j)$

At probe $p-2$: $\beta_{p-2}(i) = \sum_{j=1}^{N} a_{ij}^{IG} b_j(o_{p-1})\beta_{p-1}(j)$.

While these modifications are for the boundary when transitioning from an intergenic to gene region, respective modifications can be employed when switching from a gene to an intergenic region. If $\tilde{\alpha}_p(i)$ and $\tilde{\beta}_p(i)$ represent the modified forward and backward variables, then a modified $\tilde{\gamma}_p(i)$ (4.26) can also be computed:

$$\tilde{\gamma}_p(i) = P(q_p = S_i|O, \lambda) = \frac{\tilde{\alpha}_p(i)\tilde{\beta}_p(i)}{\sum_{i=1}^{N} \tilde{\alpha}_p(i)\tilde{\beta}_p(i)}. \tag{4.26}$$

As before, the state which maximizes $\tilde{\gamma}_p(i)$ is the estimated hidden state for probe p.

4.3.2 Modified Baum-Welch Parameter Estimates

To incorporate genomic annotation in this proposed way, the main difference in obtaining parameter estimates with the BW algorithm is the need to estimate two different sets of transition probabilities, one for the gene regions and one for the intergenic regions. The transition probabilities for genes are assumed to be constant across all genes, as are the transition probabilities for intergenic regions. First, note that the parameters π, μ, and Σ are estimated (4.27, 4.28) in the same way as before, with $\gamma_p(i)$ replaced by the modified $\tilde{\gamma}_p(i)$ (4.26):

Modified Initial Probabilities:

$$\tilde{\pi}_i = \tilde{\gamma}_1(i) \tag{4.27}$$

Modified Observation Probability Distribution Parameters:

$$\tilde{\mu_i} = \frac{\sum_{p=1}^{P} \tilde{\gamma}_p(i) o_p}{\sum_{p=1}^{P} \tilde{\gamma}_p(i)} \qquad \tilde{\Sigma_i} = \frac{\sum_{p=1}^{P} \tilde{\gamma}_p(i)(o_p - \mu_i)(o_p - \mu_i)^T}{\sum_{p=1}^{P} \tilde{\gamma}_p(i)}. \tag{4.28}$$

To obtain parameter estimates for a_{ij}^{IG} and a_{ij}^{G}, the variable $\xi_p(i,j)$ must first be modified in a similar manner as the forward and backward variables. Consider the

following changes to $\xi_p(i,j)$, moving from a intergenic to a gene region as described previously (Figure 4.7):

$$\begin{aligned}
&\text{At probe } p-2: \quad \tilde{\xi}_{p-2}(i,j) = \frac{\tilde{\alpha}_{p-2}(i)a_{ij}^{IG}b_j(o_{p-1})\tilde{\beta}_{p-1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\tilde{\alpha}_{p-2}(i)a_{ij}^{IG}b_j(o_{p-1})\tilde{\beta}_{p-1}(j)} \\
&\text{At probe } p-1: \quad \tilde{\xi}_{p-1}(i,j) = \frac{\tilde{\alpha}_{p-1}(i)(0.5*a_{ij}^{IG}+0.5*a_{ij}^{G})b_j(o_p)\tilde{\beta}_p(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\tilde{\alpha}_{p-1}(i)(0.5*a_{ij}^{IG}+0.5*a_{ij}^{G})b_j(o_p)\tilde{\beta}_p(j)} \\
&\text{At probe } p: \quad \tilde{\xi}_{p}(i,j) = \frac{\tilde{\alpha}_{p}(i)a_{ij}^{G}b_j(o_{p+1})\tilde{\beta}_{p+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\tilde{\alpha}_{p}(i)a_{ij}^{G}b_j(o_{p+1})\tilde{\beta}_{p+1}(j)}.
\end{aligned}$$

Let $\tilde{\xi}_p(i,j)$ represent these modified variables. $\tilde{\gamma}_p(i)$ and $\tilde{\xi}_p(i,j)$ can be partitioned in the following way:

Intergenic Regions: $\tilde{\gamma}_p^{ig}(i), \tilde{\xi}_p^{ig}(i,j)$ where $\{p, p+1\} \in$ Intergenic (IG) region
Gene Regions: $\tilde{\gamma}_p^{g}(i), \tilde{\xi}_p^{g}(i,j)$ where $\{p, p+1\} \in$ Gene (G) region
Boundary Regions: $\tilde{\gamma}_p^{b}(i), \tilde{\xi}_p^{b}(i,j)$ where $p \in$ IG, $p+1 \in$ G or $p \in$ G, $p+1 \in$ IG.

Note that $g = 1, ..., N_G$, $ig = 1, ..., N_{IG}$, and $b = 1, ..., N_B$ where N_G is the total number of genes in the sequence, N_{IG} is the total number of intergenic regions in the sequence, and N_B is the total number of boundary regions in the sequence. The transition probabilities for intergenic and gene regions can be estimated (4.29) as follows:

Modified Transition Probabilities:

$$\begin{aligned}
\tilde{a}_{ij}^{G} &= \frac{\sum_{g=1}^{N_G}\sum_{p=1}^{P_g-1}\tilde{\xi}_p^{g}(i,j) + 0.5 * \sum_{b=1}^{N_B}\sum_{p=1}^{P_b-1}\tilde{\xi}_p^{b}(i,j)}{\sum_{g=1}^{N_G}\sum_{p=1}^{P_g-1}\tilde{\gamma}_p^{g}(i) + 0.5 * \sum_{b=1}^{N_B}\sum_{p=1}^{P_b-1}\tilde{\gamma}_p^{b}(i)} \\
\tilde{a}_{ij}^{IG} &= \frac{\sum_{ig=1}^{N_{IG}}\sum_{p=1}^{P_{ig}-1}\tilde{\xi}_p^{ig}(i,j) + 0.5 * \sum_{b=1}^{N_B}\sum_{p=1}^{P_b-1}\tilde{\xi}_p^{b}(i,j)}{\sum_{ig=1}^{N_{IG}}\sum_{p=1}^{P_{ig}-1}\tilde{\gamma}_p^{ig}(i) + 0.5 * \sum_{b=1}^{N_B}\sum_{p=1}^{P_b-1}\tilde{\gamma}_p^{b}(i)}.
\end{aligned} \tag{4.29}$$

4.4 Summary

In this chapter, statistical methods for DNA methylation profiling experiments with tiling arrays are reviewed. In particular, hidden Markov models were introduced

as a way to model dependency between neighboring probes. While several studies have utilized HMMs for estimating the methylation status of each probe, most methods do not attempt to incorporate genomic annotation into the actual data analysis, and parameter estimation is often *ad hoc*. Evidence from previous studies indicate that using genomic annotation information in the context of a HMM framework may be a beneficial way to model the data. Here, a method is proposed in which transition probabilities are allowed to differ for genes and intergenic regions, with updates to the forward-backward and Baum-Welch algorithms to include this extra information in state and parameter estimation. The next step is to investigate whether incorporating genomic annotation information in a HMM in this manner improves DNA methylation status prediction.

5. SIMULATION STUDIES

Simulation studies are employed to investigate the importance of incorporating genomic annotation into the hidden Markov model framework for DNA methylation profiling studies. Data are simulated for a HMM with transition probabilities that differ between genes and intergenic regions. The first study assumes that HMM parameters are known and focuses on evaluating the effectiveness of using genomic annotation for methylation status estimation via comparison of the standard forward-backward (FB) algorithm (Baum et al., 1970; Baum, 1972) and the modified FB algorithm described in Section 4.3.1. In the second study, HMM parameters are estimated from the data using both the standard (Baum et al., 1970) and modified (Section 4.3.2) Baum-Welch (BW) algorithms to assess the performance of using genomic annotation when model parameters are unknown, as for real data. Additionally, states are also estimated by conducting paired t-tests for each probe, as well as implementing the forward-backward algorithm with *ad hoc* HMM parameter estimates. Comparison of these results to those using the BW algorithm investigates the value of modeling dependence between neighboring probes and obtaining maximum likelihood estimates for HMM model parameters. A common set of simulation settings is used for both studies and are described in the next section.

5.1 Simulation Settings

For a hidden Markov model, a set of hidden states and observations given the model parameters $\lambda=(A,B,\pi)$ can be simulated according to the method described in section 4.2.1. Hidden states (0=Unmethylated, 1=Methylated) and observations are simulated for a genomic region of 2000 probes covering 20 genes (Figure 5.1), with 58 probes representing each gene. The total number of probes is similar in magnitude

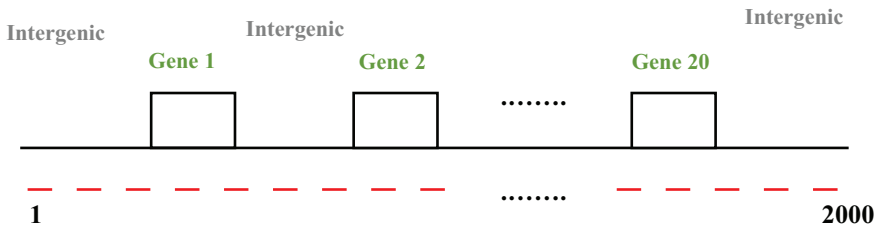

Figure 5.1. A genomic region with 2000 probes and 20 genes.

to that of the *Arabidopsis thaliana* chromosome 4 tiling array custom-designed by Lippman et al. (2004), representing a small-scale study for the purpose of evaluating model performance. The density of genes per total number of probes and probes per gene reflects those averages for probes on the *Arabidopsis* Affymetrix® whole genome tiling array.

Two different DNA methylation patterns are simulated with transition probabilities for genes and intergenic regions given in Table 5.1. The initial state distribution is assumed to be $\pi = (0.5, 0.5)$ for both patterns. At boundaries where an intergenic region changes to a gene region (or vice versa), the hidden state of the first probe in a new region is simulated from the average of the intergenic and gene transition probabilities: $a_{ij}^{B} = 0.5 * a_{ij}^{IG} + 0.5 * a_{ij}^{G}$. The first DNA methylation pattern is representative of mosaic DNA methylation as described in Suzuki and Bird (2008) where regions of dense methylation are interspersed with less dense, more variable DNA methylation regions. In particular, for organisms such as *Arabidopsis thaliana* genes are often either densely methylated or not methylated at all, whereas intergenic regions may be more variable. This is reflected by the high transition probability ($a_{ii}^{G} = 0.99$) of staying in the same state for sequential probes in a gene region and a lower such probability ($a_{ii}^{IG} = 0.7$) in intergenic regions (Table 5.1). The second DNA methylation pattern assumes the transition probabilities are constant across the whole region. This pattern is used to determine whether there is a difference in model performance by using the modified versions of the forward-backward and Baum-Welch algorithms, even if the transition probabilities for genes and intergenic regions are truly the same.

Table 5.1
Model parameter settings for intergenic (a_{ij}^{IG}) and gene (a_{ij}^{G}) transition probabilities for two different DNA methylation patterns. Note that the hidden state of the first probe in a new region at a boundary of a gene and intergenic region is simulated from the average of the two transition probabilities: $a_{ij}^{B} = 0.5 * a_{ij}^{IG} + 0.5 * a_{ij}^{G}$.

DNA Methylation Pattern	a_{ij}^{IG}	a_{ij}^{G}
1: Different for Intergenic/Genes	$\begin{pmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{pmatrix}$	$\begin{pmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \end{pmatrix}$
2: Same Across Whole Region	$\begin{pmatrix} 0.85 & 0.15 \\ 0.15 & 0.85 \end{pmatrix}$	$\begin{pmatrix} 0.85 & 0.15 \\ 0.15 & 0.85 \end{pmatrix}$

Observed data are generated from a variety of different parameter settings of the following observation probability distribution:

$$\begin{aligned} b_0 \begin{pmatrix} y_{1pk} \\ y_{2pk} \end{pmatrix} &\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix} \right) \quad \text{Unmethylated Probes} \\ b_1 \begin{pmatrix} y_{1pk} \\ y_{2pk} \end{pmatrix} &\sim N\left(\begin{pmatrix} \mu_{11} \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix} \right) \quad \text{Methylated Probes.} \end{aligned} \tag{5.1}$$

These parameter settings are chosen so that in the unmethylated case, the untreated ($i = 1$) and treated ($i = 2$) means are equal and centered at zero, but in the methylated case there is a difference of (μ_{11}) between the two means. Decreasing the magnitude of μ_{11} and increasing σ 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 ρ 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 distri-

bution parameter settings (5.2) are employed to simulate three biological replicates for both DNA methylation patterns (Table 5.1) and the 2000 probes:

$$\mu_{11} = \{0.75, 1, 2\} \quad \sigma = \{1, 2\} \quad \rho = \{0.3, 0.7\}. \tag{5.2}$$

Averages over the 3 biological replicates are calculated for input into the forward-backward or the Baum-Welch algorithms, with modified observation probability distribution given by Equation 4.15. These data are simulated 1000 times.

5.2 Simulation Study 1: Investigating Importance of Genomic Annotation

5.2.1 Study Goal and Model Comparison

In this simulation study, the goal is to compare the performance of a HMM which incorporates genomic annotation into hidden state estimation and a HMM model that does not utilize this information. Models are evaluated under the best case scenario that the model parameters are known, and thus no parameter estimation is required. Simulated observed data and the true model parameters are used to estimate hidden states with both the standard (Baum et al., 1970; Baum, 1972) and the modified (Section 4.3.1) forward-backward algorithms. The following two models are compared.

Unannotated Model: Genomic annotation information is ignored and the transition probabilities are assumed to be the same across the entire genomic region. For the model with truly different intergenic and gene transition probabilities (Table 5.1), the common transition probabilities are assumed to be the weighted average of a_{ij}^{IG} and a_{ij}^{G} (i.e., $a_{ij} = \begin{pmatrix} 0.87 & 0.13 \\ 0.13 & 0.87 \end{pmatrix}$). The standard forward-backward algorithm (Baum et al., 1970; Baum, 1972) is employed for hidden state estimation.

Annotated Model: Genomic annotation information is incorporated into the HMM by assuming the gene has different transition probabilities (a_{ij}^{G}) than the inter-

genic region (a_{ij}^{IG}) (Table 5.1). The modified forward-backward algorithm (Section 4.3.1) which integrates these potential 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. Note that when the transition probabilities are truly constant across the region, the modified forward-backward algorithm (Annotated Model) should be equivalent to the standard (Unannotated Model) forward-backward algorithm when given the true parameters. Differences in the two models should be apparent when the transition probabilities are truly different for genes and intergenic regions.

5.2.2 Results and Conclusions

Figure 5.2 illustrates the results for the first DNA methylation pattern (Table 5.1) with different transition probabilities for genes and intergenic regions. The proportion of correctly predicted states is given for both the annotated and the unannotated models for each of the observation probability distribution parameter settings (5.2). Across all settings, the annotated model outperforms the unannotated model. The difference in the magnitude of performance between the two models increases as σ increases and the mean difference between the untreated and treated samples (μ_{11}) decreases, meaning that the annotated model can more accurately model noisy data and detect smaller mean differences. The difference in model performance also appears to slightly increase as the correlation between the samples taken on the same subject (ρ) decreases, but this change in performance is only slight. Figure 5.3 shows the results for the second DNA methylation pattern (Table 5.1) with the same set of transition probabilities across the whole genomic region. As expected, for all of the observation probability distribution settings, the unannotated and the annotated model perform identically. The results from this simulation study show that the an-

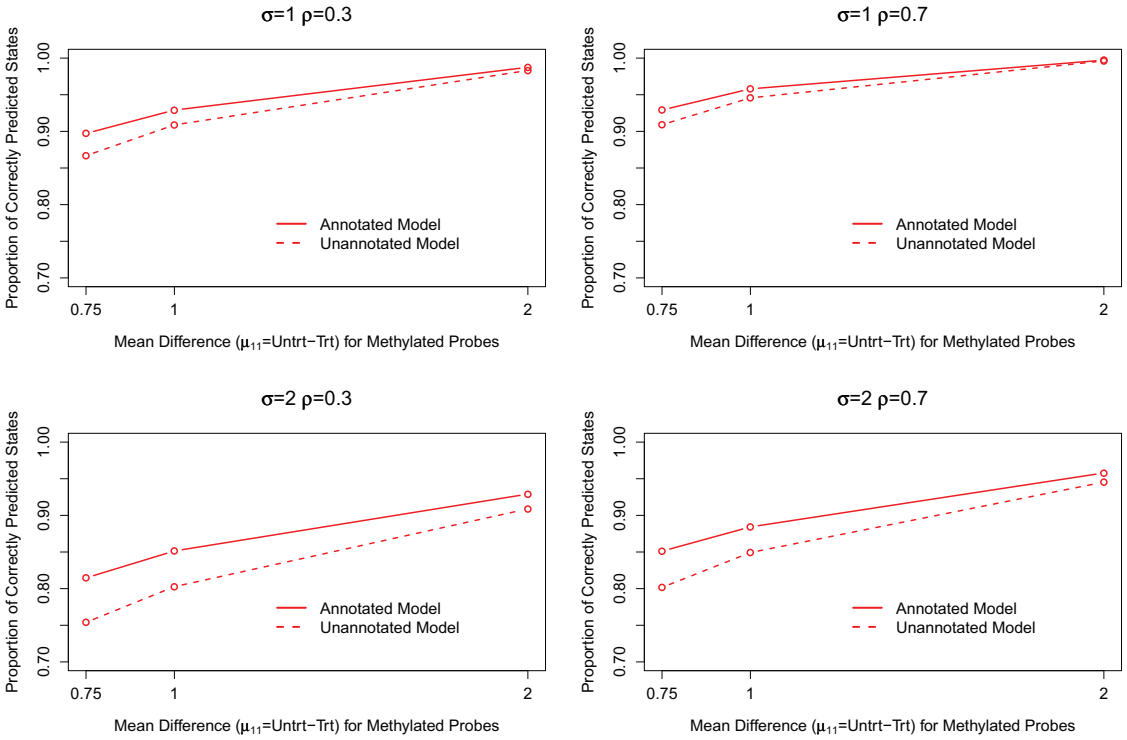

Figure 5.2. Results from Simulation Study 1 for the first DNA methylation pattern (Table 5.1) when transition probabilities are different for genes and intergenic regions. The proportion of states predicted correctly for the annotated and unannotated models is plotted for each of the μ_{11} parameter settings of the observation probability distribution (5.2). Separate plots are shown for each combination of the σ and ρ parameters of the observation probability distribution.

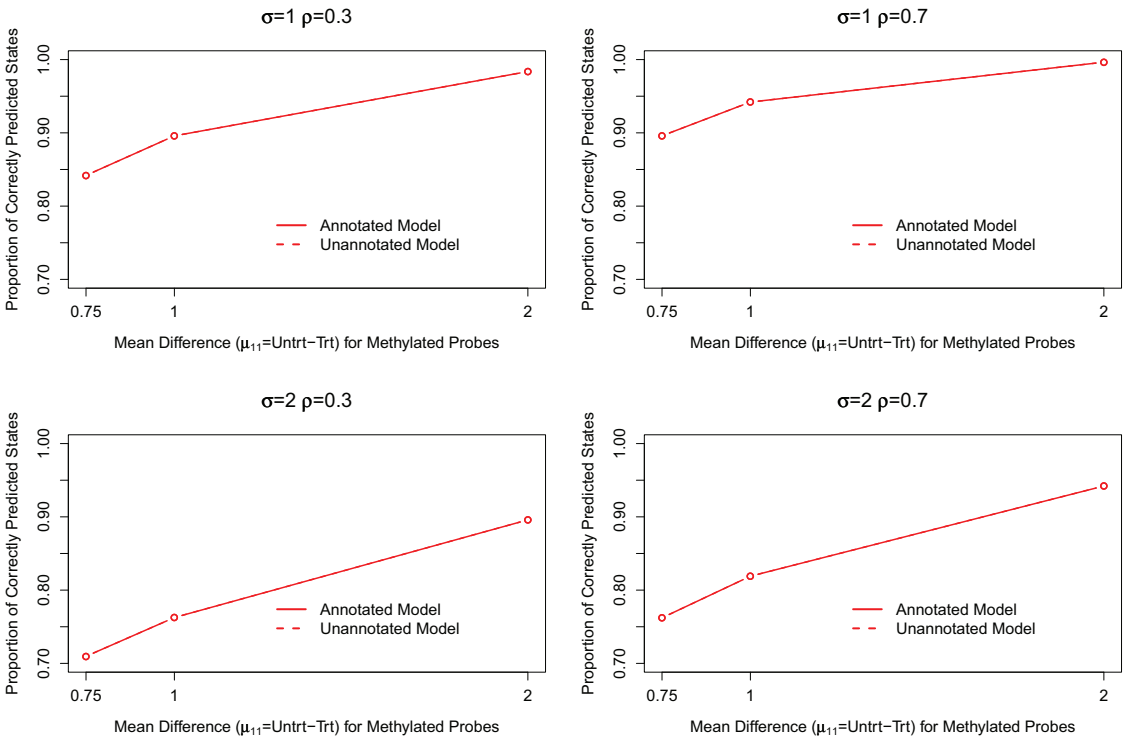

Figure 5.3. Results from Simulation Study 1 for the second DNA methylation pattern (Table 5.1) when transition probabilities are constant across the whole region. The proportion of states predicted correctly for the annotated and unannotated models is plotted for each of the μ_{11} parameter settings of the observation probability distribution (5.2). Separate plots are shown for each combination of the σ and ρ parameters of the observation probability distribution. Note that the performance of the two models is identical, resulting in the appearance of only one line.

notated model performs better than the unannotated when there truly is a difference in transition probabilities for genes and intergenic regions, and that using the annotated model does not decrease model performance if the transition probabilities are the same across the whole region.

5.3 Simulation Study 2: Investigating Parameter Estimation with Genomic Annotation

5.3.1 Study Goal and Model Comparison

In the previous simulation, true model parameters are used to estimate the hidden state sequence via the FB algorithm. However, in reality, the true model parameters are not known and must be estimated. The goal of this simulation study is to investigate the performance of the modified Baum-Welch (described in Section 4.3.2) algorithm which incorporates genomic annotation information and yields separate transition probability estimates for intergenic and gene regions. This model will be compared to the standard BW (Baum et al., 1970) algorithm, *ad hoc* parameter estimation, and paired t-tests to examine the effects of incorporating annotation, obtaining maximum likelihood estimates, and modeling dependency between probes, respectively. The following six models are compared.

Independent Paired t-tests: For each probe, a paired t-test (with $\mu_0 = 0$) is performed as described in Section 4.1.1 to determine whether the probe is methylated or not. Dependence between probes and genomic annotation are both ignored in this model.

FB with Initial Estimates: The estimated methylation status from the paired t-tests can be used to empirically estimate the HMM model parameters. These initial *ad hoc* estimates are then used with the standard (**Unannotated**) and modified (**Annotated**) FB algorithms to estimate the methylation status of each probe. While

these initial estimates do not maximize the likelihood $P(O|\lambda)$, this method does incorporate dependence between probes via the FB algorithm.

FB with BW Estimates: The initial parameter estimates calculated from paired t-test results can be used as starting values in the standard (**Unannotated**) and modified (**Annotated**) BW algorithms to obtain updated parameter estimates which locally maximize $P(O|\lambda)$. These updated parameter estimates can then be used with appropriate FB algorithm to estimate the methylation status of each probe.

FB with True Parameters: Since the true parameter values are known, the FB algorithm can be employed to estimate the methylation states given the true model parameters. This is used as a basis for comparison to the other models, as it should yield the best possible results.

Each of these models will yield estimated methylation states for all probes. The estimated states are again compared to the true methylation status by determining the proportion of states predicted correctly and averaging across the 1000 data sets. Note that, unlike the forward-backward algorithm, when the transition probabilities are truly constant across the region, the modified Baum-Welch algorithm (Annotated Model) may not yield the same results as the standard (Unannotated Model) Baum-Welch algorithm since separate transition probabilities for the gene and intergenic region are estimated from the data. Results are compared to determine how the modified BW affects model performance if genomic annotation is not needed. Differences between all six models should be apparent when the transition probabilities are truly different for genes and intergenic regions.

5.3.2 Results and Conclusions

The results for the first DNA methylation pattern (Table 5.1) when transition probabilities are different for genes and intergenic regions are shown in Figure 5.4. As expected, the model which utilizes the true parameter estimates to estimate hidden states performs the best across all parameter settings. The paired t-tests and FB with *ad hoc* parameter estimation perform poorly across all parameter settings, even though the FB with *ad hoc* parameters performs better or similar to the paired t-tests in most cases. Parameter estimation with Baum-Welch prior to hidden state estimation with the FB algorithm greatly improves model performance. In particular, when genomic annotation is incorporated, this model often achieves the same performance as if the true parameters were known. Similar to Simulation Study 1, it is clear that all models perform the worst when μ_{11} and ρ are the smallest, and σ is the largest. Similar trends are seen in the results for the second DNA methylation pattern (Table 5.1), which assumes a constant transition probability across the whole region (Figure 5.5). The main difference being that performance of the unannotated and the annotated models is the same across all observation probability distribution settings. The results from this simulation study further indicate that incorporating genomic annotation into the HMM framework is beneficial for hidden state estimation when there truly is a difference in transition probabilities for genes and intergenic regions. It also demonstrates that the model which incorporates genomic annotation will not affect the proportion of correctly predicted states if the transition probabilities are truly the same across the whole region. Additionally, modeling the dependence between probes improves model performance the most when the Baum-Welch algorithm is employed to obtain maximum likelihood estimates.

5.4 Summary

Genomic annotation is an important source of biological information that is often used with DNA methylation profiling studies to better understand the distribution

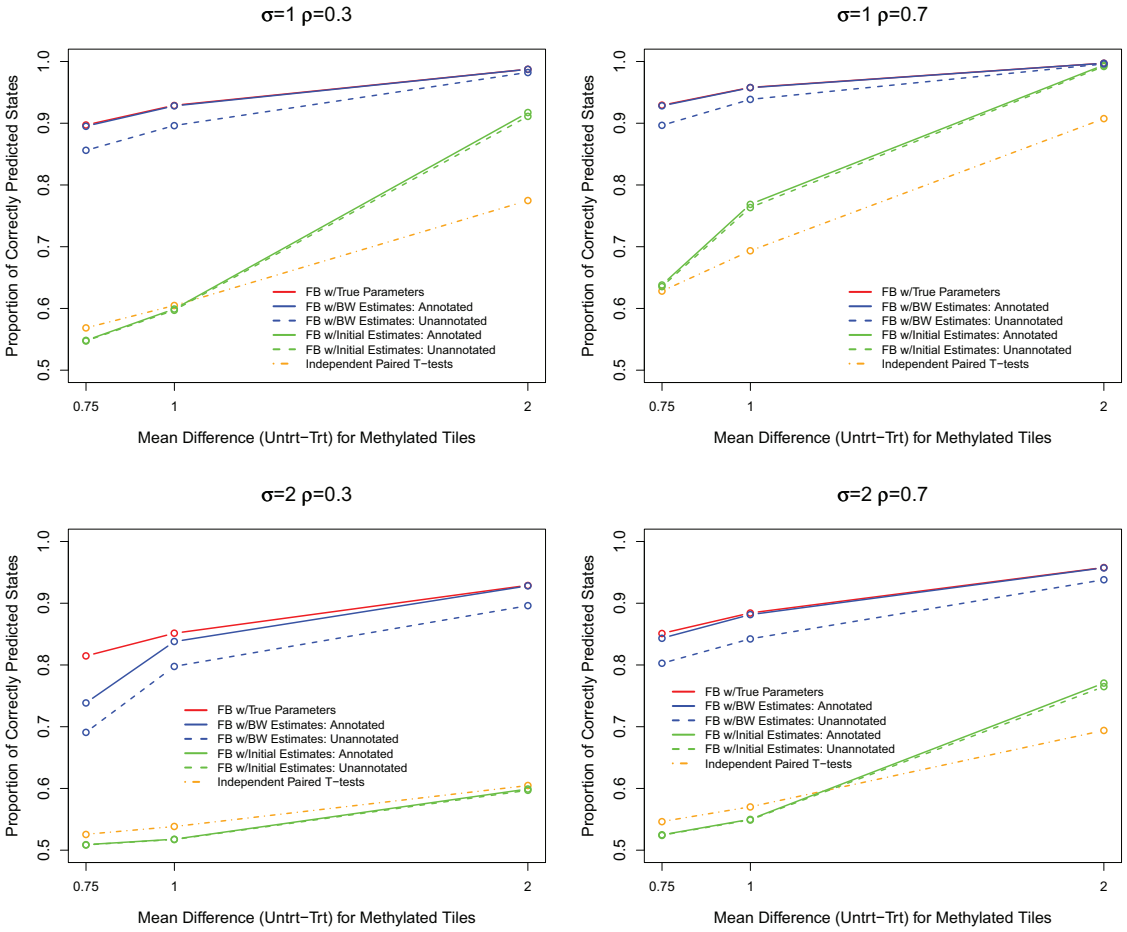

Figure 5.4. Results from Simulation Study 2 for the first DNA methylation pattern (Table 5.1) when transition probabilities differ for genes an intergenic regions. Proportion of states predicted correctly for each of the six models (indicated by different colors and line types) are shown for each of the μ_{11} parameter settings of the observation probability distribution (5.2). Separate plots are shown for each combination of the σ and ρ parameters of the observation probability distribution.

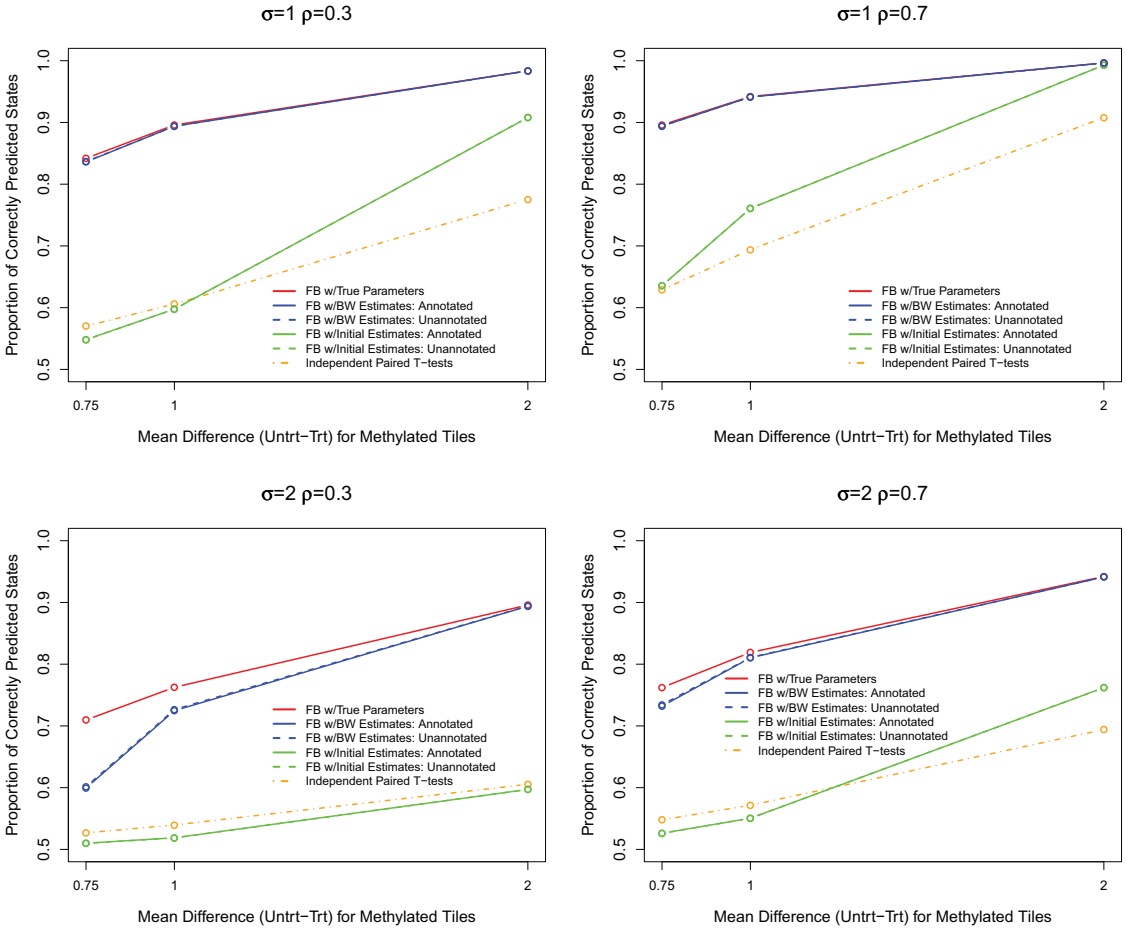

Figure 5.5. Results from Simulation Study 2 for the second DNA methylation pattern (Table 5.1) when transition probabilities are constant across the whole region. Proportion of states predicted correctly for each of the six models (indicated by different colors and line types) are shown for each of the μ_{11} parameter settings of the observation probability distribution (5.2). Separate plots are shown for each combination of the σ and ρ parameters of the observation probability distribution. Note that the performance of the annotated and unannotated models for the FB with BW estimates is nearly identical, resulting in the appearance of only one line. The same is true for the annotated and unannotated models for the FB with initial estimates.

of this epigenetic modification across the genome. Typically, genomic annotation is utilized after the completion of a statistical analysis to explore DNA methylation patterns for different genomic elements. While this approach has provided valuable insight in the study of DNA methylation, it may be beneficial to incorporate information from genomic annotation into a statistical analysis. In chapter 4, a method for incorporating genomic annotation information into a hidden Markov model framework was proposed by allowing transition probabilities to differ between genes and intergenic regions. Modified versions of the forward-backward and Baum-Welch algorithm were introduced to incorporate these differences in transition probabilities into standard state and parameter estimation algorithms.

In this chapter, the proposed method for incorporating genomic annotation information into a HMM framework for analysis of DNA methylation tiling array data is investigated via two simulation studies. Results from the simulation studies show that incorporating genome annotation into HMMs improves prediction of DNA methylation status if there truly are differences in transition probabilities between intergenic and gene regions. In this case, the modified FB and BW algorithms perform better than the standard FB and BW algorithms. More specifically, the algorithms which incorporate genomic annotation are able to more accurately predict hidden states, even when the observations are noisy and the mean difference between the untreated and treated sample for methylated probes is small. In addition to incorporating genomic annotation, the importance of BW parameter estimation in correctly estimating hidden states is demonstrated by its notable improvement in model performance compared to employing *ad hoc* parameter estimates with the FB algorithm or conducting paired t-tests, which ignore probe dependency and genomic annotation. Finally, in the case that transition probabilities are truly constant across the region, the performance of the annotated model is the same as that of the unannotated model. Ultimately, these results indicate that using the modified FB and BW algorithms which incorporate genomic annotation improves model performance if ge-

nomic annotation information is needed, and can be used even when this information is not needed without an affect on model performance.

6. APPLICATION TO REAL DATA: DNA METHYLATION PROFILING IN *ARABIDOPSIS THALIANA*

The method proposed in this research for incorporating genomic annotation information into a HMM framework for DNA methylation profiling studies (Chapter 4) is applied to two real data sets from previously published work. Both studies use tiling arrays to investigate DNA methylation in the model plant *Arabidopsis thaliana*. The first study (Lippman et al., 2004) utilizes a custom-designed tiling array, whose design is described in Section 2.3.1, to investigate DNA methylation in the heterochromatic knob (*hk4S*) on the short arm of chromosome 4. This region is known to contain many transposable elements and repetitive DNA (Martienssen and Colot, 2001), which are often heavily methylated. In that work, an ANOVA model (4.1) was employed to determine statistical significance of each probe. The second study (Zhang et al., 2006) is the first genome-wide DNA methylation profiling experiment for *Arabidopsis thaliana*. The Affymetrix® whole genome tiling array, described in Section 2.3.2, is employed to accomplish this task and the HMM proposed by Ji and Wong (2005) is used to determine the DNA methylation status of all probes. Here, genomic annotation is incorporated into the analysis of these two data sets via application of the modified forward-backward and Baum-Welch algorithms described in Section 4.3.1 and 4.3.2. Results are compared to those from the methods employed in each of the studies. Additionally, results for the heterochromatic knob region of chromosome 4 are compared between the two studies.

6.1 Chromosome 4 Tiling Array Data

6.1.1 Description of Lippman et al. (2004) Study

Before whole genome tiling arrays became available, Lippman et al. (2004) designed a cDNA tiling array to conduct a small scale study of epigenetic modifications in the heterochromatic knob (*hk4S*) region of chromosome 4. This tiling array covers base pair positions 1,201,322 to 2,673,088 on chromosome 4, including the knob which is located between base pair positions 1,600,000 to 2,330,000. 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. Since the region contains numerous transposable elements and repetitive DNA, heavy methylation is expected inside the region. Details about the design and genomic annotation of the Lippman et al. (2004) custom-designed tiling array are given in Section 2.3.1. Briefly, recall that there are 1407 unique probes (each replicated two to four times) represented on the array that cover the 1.5 megabase (Mb) region centered on *hk4S*. On average, probes are 995 base pairs in length with the possibility of gaps or overlaps between probes. Of the 1407 probes, 71.6% of them lie in gene regions, with an average of three probes per gene. A set of 680 probes located in the euchromatic region outside the knob can be used as controls, since they are known to be unmethylated.

For the purposes of this research, 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. To study DNA methylation, Lippman et al. (2004) employ the use of a methylation restriction enzyme, McrBC, to separate methylated and unmethylated DNA. In this technique, a DNA sample collected from an individual is split into two subsamples. One of the subsamples (treated sample) is digested with McrBC to remove sequences which contain DNA methylation. The other sample is left untreated and is representative of total DNA, with both unmethylated and methylated DNA retained. A dye swap is then performed

by splitting the treated and untreated samples into two subsamples, labeling each of the subsamples with a different dye, and hybridizing treated and untreated samples with different dye labels to the custom-designed tiling arrays. 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 ANOVA model (4.1) and hypotheses tests (4.2) described in Section 4.1.1. Yoo (2008) later reanalyzed these data by conducting the updated set of hypotheses tests (4.3), which address issues specific to DNA methylation tiling array data. 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 hidden Markov model proposed in this research which incorporates genomic annotation information. The HMM which uses the standard forward-backward algorithm, as well as the Baum-Welch algorithm, without genomic annotation information is also applied for comparison purposes. Results from the two HMMs and the ANOVA employed by Yoo (2008) are compared. The same background correction and normalization process is performed on the data prior to all analyses.

Also, note that the starting point and direction of estimation for HMMs have the potential to affect model performance. While data in simulation studies are generated according to a certain direction, the direction is less clear for real DNA methylation profiling data. Recall that the DNA sequence consists of two strands in opposing directions and genes can be located on either strand. Thus, it is unclear which direction the HMM estimation should be applied since there is not one clear direction to the DNA sequence. To investigate the impact of the estimation direction, the HMM with genomic annotation information is run in both directions and results for the two directions are compared.

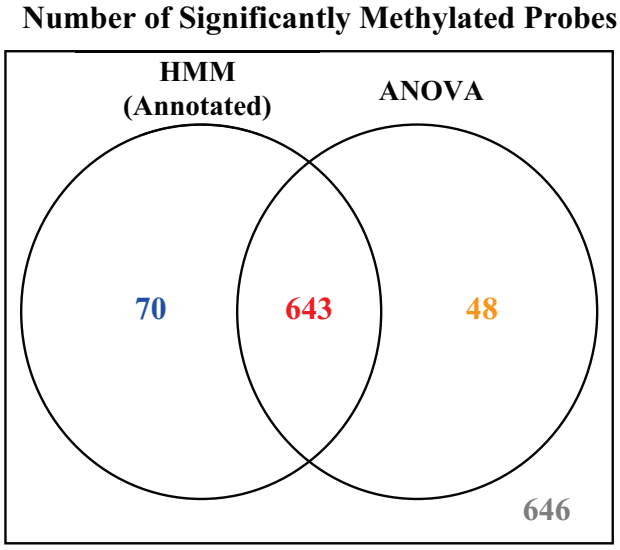

Figure 6.1. Venn diagram comparing the number of significantly methylated probes identified by the HMM model with genomic annotation and the ANOVA model. Both methods find 643 methylated probes and 646 unmethylated probes. The ANOVA model identifies 48 methylated probes that the HMM with annotation does not; however, the HMM with annotation identifies 70 probes as being methylated that the ANOVA does not.

6.1.2 Comparison of Results

Figure 6.1 gives a comparison of the number of significantly methylated probes identified by the ANOVA model and the HMM model, which incorporates genomic annotation. Note that the direction of estimation does not impact the annotated HMM results in these data, as the estimated DNA methylation status in both directions is identical for 99.9% of the probes. Thus, results for only one direction are presented here. Of the 1407 probes represented on the array, both methods identify 643 of these as being significantly methylated. The ANOVA model identifies 48 probes that the HMM with genomic annotation does not, yielding a total of 49.1% of probes on the array that are significantly methylated. The HMM with genomic annotation identifies a total of 50.7% methylated probes, and 70 of these are not significantly methylated in the ANOVA method. Of probes located in the heterochromatic knob

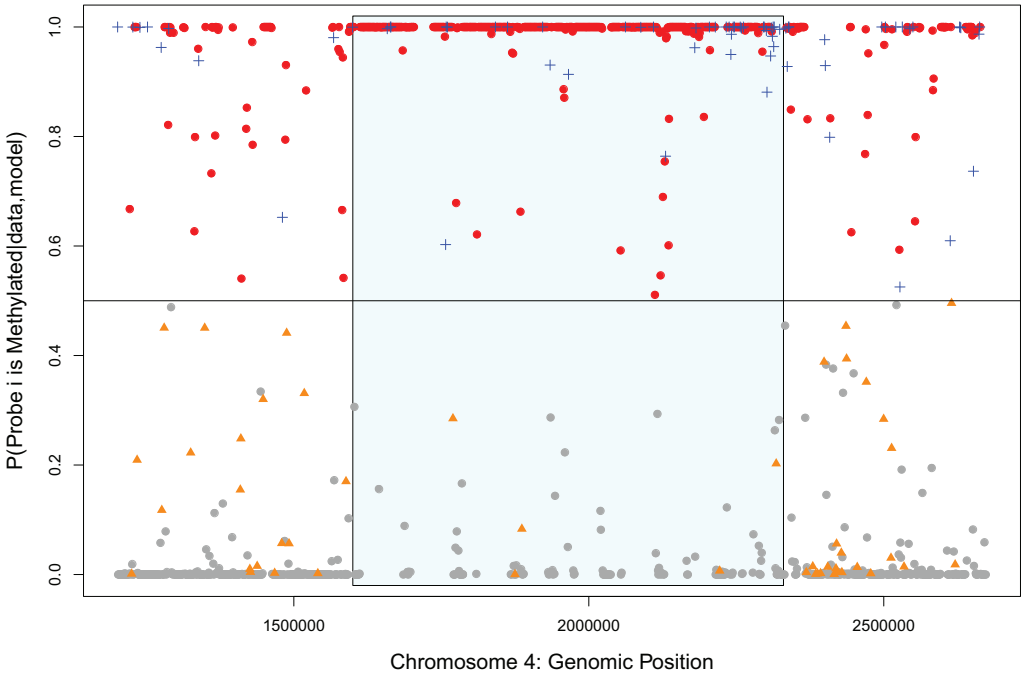

Figure 6.2. The probability of each probe being in the methylated state (given the model parameters and data) plotted by the genomic location of each probe's start position. The colors of the symbols correspond to the colors in the Venn diagram (Figure 6.1), where red points (dots) are probes that are significantly methylated using both methods, blue points (crosses) are only found methylated in the HMM with annotation, 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).

region, 74.9% are significantly methylated using ANOVA and 79.9% are identified as methylated using the HMM with genomic annotation. The high percentage of DNA methylation found in the knob region by both methods 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 6.2 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 HMM with annotation 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 HMM with annotation (orange points) were mostly located in the euchromatic region outside the knob. On the other hand, the HMM with annotation identified several methylated probes at the right end of the heterochromatic knob that the ANOVA model did not (blue points). To estimate the false positive rate, the percentage of control probes (which are known to be unmethylated) that are identified as significantly methylated by both methods is calculated. For the ANOVA model, 24.4% of control probes are significantly methylated. This is similar (but slightly better) for the HMM with annotation, which identifies 21.9% control probes as being methylated.

To evaluate the effect of incorporating genomic annotation in the model, data are analyzed with both the annotated and unannotated HMMs. For these data, there are no differences in DNA methylation status estimates between the the two models. The parameter estimates for the annotated model are given below:

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

$$\hat{\mu}_0 = \begin{pmatrix} 0.1147 \\ 0.4338 \end{pmatrix} \qquad \hat{\mu}_1 = \begin{pmatrix} 0.5547 \\ -0.9993 \end{pmatrix}$$

$$\hat{\sigma}_{01} = 0.6867, \hat{\sigma}_{02} = 0.6108, \hat{\rho}_0 = 0.8100 \qquad \hat{\sigma}_{11} = 1.5107, \hat{\sigma}_{12} = 0.9094, \hat{\rho}_0 = 0.8343.$$

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 ~ 0.86). Although this analysis did not reveal evidence for

a difference in transition probabilities for genes and intergenic regions, employing the annotated HMM allows the direct investigation of such patterns through statistical analysis that is not possible by previous methods. Also, it appears that there is more variation among the methylated probes ($\hat{\sigma}_{11}$,$\hat{\sigma}_{12}$) than the unmethylated probes ($\hat{\sigma}_{01}$,$\hat{\sigma}_{02}$). For both unmethylated and methylated probes, the correlation between samples collected on the same individual is fairly high (~ 0.8). Note that for this tiling array design, there are only an average of three probes per gene due to longer the probe length. Many genes and intergenic regions only contain one probe. In this case, a single probe in a region is considered to be a boundary probe, since no within region transition occurs.

6.2 Whole Genome Affymetrix® Tiling Array Data

6.2.1 Description of Zhang et al. (2006) Study

The first genome-wide DNA methylation profiling study was conducted by Zhang et al. (2006) for *Arabidopsis thaliana* using Affymetrix® whole genome tiling arrays. Details about the design of this tiling array are described in Section 2.3.2. Recall that there are ∼3 million probes that cover non-repetitive regions of all five *Arabidopsis* chromosomes. Each of these probes is 25 bases in length with an average gap of 10 bases between probes. Genes are represented by 59.7% of the probes, with 58 probes on average per gene. Zhang et al. (2006) investigate genome-wide DNA methylation in wild-type Columbia and two *Arabidopsis* mutants (*met1* and *drm1 drm2 cmt3*) through a methylcytosine immunoprecipitation technique that is employed to separate unmethylated and methylated DNA. The separate samples of unmethylated and methylated DNA collected from the same individual are each hybridized to an Affymetrix® whole genome tiling array. In this case, the intensity values of truly methylated probes are still expected to be greater in the methylated sample than in the unmethylated sample. However, the unmethylated sample should yield higher values than the methylated sample when the probes are truly unmethylated (rather

than having a similar intensity as was the case for McrBC, where unmethylated DNA is retained in both samples). Three biological replicates are studied, yielding a total of six tiling arrays.

To determine the methylation status of each probe represented on the tiling array, Zhang et al. (2006) utilize Tilemap software which employs the HMM developed by Ji and Wong (2005). For the purposes of this research, the wild type Columbia *Arabidopsis* samples are reanalyzed with the HMM developed here to incorporate genomic annotation to better understand naturally occurring DNA methylation and allow comparison to the Lippman et al. (2004) wild-type Columbia results from the heterochromatic knob of chromosome 4. Here, results between the annotated HMM and Tilemap, as well as between the annotated and unannotated HMMs are compared. Note that the pre-processing steps of background correction and normalization are both conducted in the same manner for all analyses. Also, HMM estimation is performed in both directions and results compared.

6.2.2 Comparison of Results

Figure 6.3 shows the differences in DNA methylation status results between the HMM which incorporates genomic annotation and the HMM used in Tilemap. As in the Lippman et al. (2004) data, the direction of HMM estimation does not heavily impact the annotated HMM results, with 99.9% of probes having the same estimated DNA methylation status in both directions. Again, results from only one direction are presented here due to this high agreement. It is evident that the annotated HMM identifies over twice as many methylated probes as Tilemap. Overall, 27.51% of probes are identified as being methylated using the HMM with genomic annotation compared to 9.78% of probes using Tilemap. Both of these methods offer a similar breakdown in terms of methylated probes in genes, with 68.36% and 70.43% of significantly methylated probes using Tilemap and the HMM with annotation, respectively,

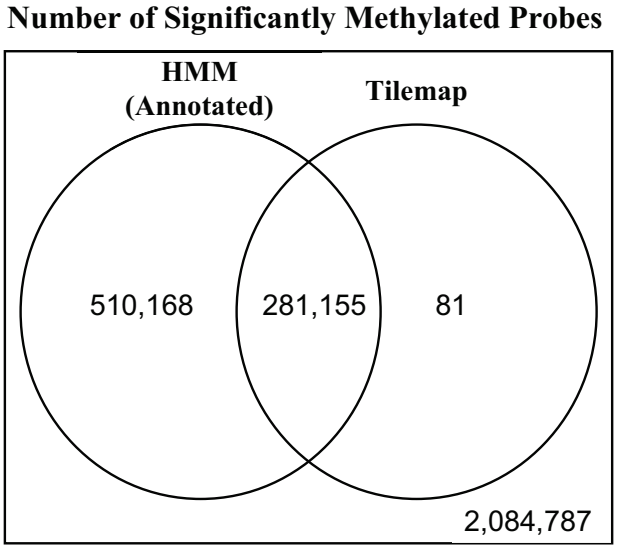

Figure 6.3. Venn diagram comparing the number of significantly methylated probes identified by the HMM model with genomic annotation and the HMM with Tilemap. Both methods find 281,155 methylated probes and 2,084,787 unmethylated probes. The HMM model with genomic annotation identifies many more methylated probes (510,168) that the HMM with Tilemap does not; whereas the HMM with Tilemap only identifies 81 probes as being methylated that the HMM with annotation does not.

occurring within genes. Of the probes identified as being methylated by the HMM with annotation, but not by Tilemap, 71.56% of them occur within genes.

To further examine the differences between the HMM with annotation and Tilemap, the distribution of probabilities calculated from the annotated HMM of probes being methylated, given the model and data, is investigated (Figure 6.4). For probes that are identified as methylated using both methods (Figure 6.4, upper left), it is clear that the probability of being methylated is high and is greater than 0.90 for 99.7% of the probes. Similarly, for probes that are identified as unmethylated using both methods (Figure 6.4, upper right), the probability of being methylated is low and below 0.10 for 91.8% of the probes. However, results for probes identified as methylated by only one of the methods are more variable. Methylated probes using the annotated HMM only (Figure 6.4, lower left) still exhibit a very high probability of

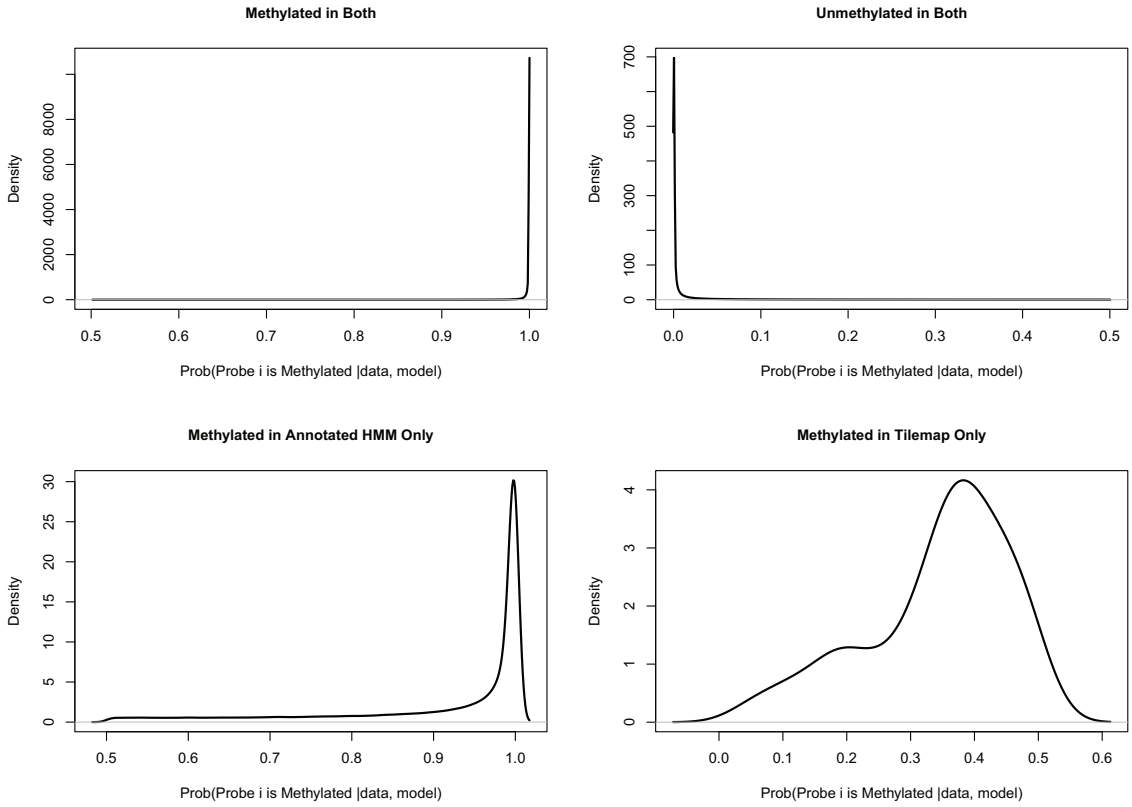

Figure 6.4. Density plots for the probability of each probe being in the methylated state (given the model parameters and data) calculated from the annotated HMM. Separate plots are given for probes that are identified as methylated using both the annotated HMM and Tilemap (upper left), unmethylated using both methods (upper right), methylated with the annotated HMM only (lower right), and methylated using Tilemap only (lower right). Note that all probabilities for probes identified as methylated using both methods and the annotated HMM only are above 0.5, while the probabilities for probes identified as unmethylated in both methods and methylated using Tilemap only (i.e. unmethylated with the annotated HMM) are below 0.5.

being methylated (greater than 0.90 for 72.6% of the probes). For probes identified as methylated by Tilemap only (Figure 6.4, lower right), the probability of being methylated using the annotated HMM probabilities is below 0.5 since the probes are

unmethylated using the annotated HMM. However, the results for probes identified as methylated by Tilemap only are much more variable and closer to 0.5 than in any of the other graphs, with an average probability of being methylated of 0.34.

These results indicate that probes identified as methylated or unmethylated by both methods show strong evidence for their predicted DNA methylation status. Probes identified as methylated by the annotated HMM only also seem to show a high degree of evidence for their predicted methylation status, while probes identified as methylated by Tilemap only seem more questionable. It is difficult to determine the exact reason for the differences in results between these to methods, since in addition to the extra information provided by genomic annotation in the annotated HMM, different summary statistics are used on the observations and parameter estimation is performed in different ways. Tilemap utilizes an empirical Bayes t-statistic (Ji and Wong, 2005), while the annotated HMM takes the average of the biological replicates to summarize the data. The annotated HMM assumes the observations follow a bivariate normal distribution and employs the Baum-Welch algorithm for estimation of the HMM model parameters. Tilemap does not assume a specific distribution for the observations and uses a method called unbalanced mixture subtraction (UMS) to determine the initial probabilities and the observation probability distributions. The transition probabilities are determined in an *ad hoc* manner by using prior knowledge about the typical length of a methylated region (Ji and Wong, 2005). Additional simulation studies could help to better understand the performance difference in these two models. However, for these data, it is evident that the annotated HMM identifies many more methylated probes than Tilemap, and these probes have a high probability of being methylated according to the model.

Figure 6.5 shows results for the HMM model which uses genomic annotation via the modified FB and BW algorithms and the unannotated model which uses the standard versions of the algorithms. Although the results are fairly similar, the annotated HMM identifies more methylated probes than the unannotated model. Of these probes that are not found to be methylated in the unannotated model, 78.55%

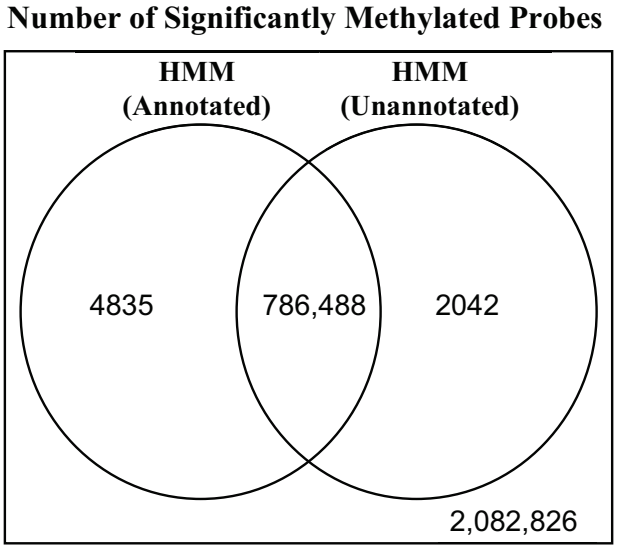

Figure 6.5. Venn diagram comparing the number of significantly methylated probes identified by the annotated and unannotated HMM models. Both methods find 786,488 methylated probes and 2,082,826 unmethylated probes. The HMM model with genomic annotation identifies 4835 methylated probes that the unannotated HMM does not; whereas the HMM without annotation identifies 2042 probes as being methylated that the HMM with annotation does not.

of them are located in genes. Of the significantly methylated probes found with the unannotated model, but not the annotated model, only 49.02% of those probes are located in genes.

Table 6.1 shows the estimates for the transition probabilities from the annotated HMM model. Note that if a region only contains one probe, it is considered a boundary probe. As in the Lippman et al. (2004) data, the transition probabilities are similar between the intergenic and gene regions, with some small differences that likely lead to the unannotated and annotated HMM models detecting some differences in methylated probes, as shown in Figure 6.5. Utilizing the annotated model is the only way to allow these transition probabilities to be compared. The parameter estimates of the observation probability distribution are given in Table 6.2. Note that, as expected, the means for the unmethylated probes ($\hat{\mu}_0$) reveal that the sam-

Table 6.1
Transition probability estimates from the modified BW algorithm for the HMM which incorporates genomic annotation. The left column gives estimates for intergenic regions and the right column gives estimates for genes. Results are given for each of the five *Arabidopsis* chromosomes.

Chr.	$\hat{a}_{ij}^{IG}$	$\hat{a}_{ij}^{G}$
1	$\begin{pmatrix} 0.9891 & 0.0109 \\ 0.0387 & 0.9613 \end{pmatrix}$	$\begin{pmatrix} 0.9795 & 0.0205 \\ 0.0428 & 0.9572 \end{pmatrix}$
2	$\begin{pmatrix} 0.9858 & 0.0142 \\ 0.0344 & 0.9656 \end{pmatrix}$	$\begin{pmatrix} 0.9821 & 0.0179 \\ 0.0339 & 0.9661 \end{pmatrix}$
3	$\begin{pmatrix} 0.9882 & 0.0118 \\ 0.0311 & 0.9689 \end{pmatrix}$	$\begin{pmatrix} 0.9823 & 0.0177 \\ 0.0365 & 0.9635 \end{pmatrix}$
4	$\begin{pmatrix} 0.9870 & 0.0130 \\ 0.0332 & 0.9668 \end{pmatrix}$	$\begin{pmatrix} 0.9798 & 0.0202 \\ 0.0388 & 0.9612 \end{pmatrix}$
5	$\begin{pmatrix} 0.9877 & 0.0123 \\ 0.0400 & 0.9600 \end{pmatrix}$	$\begin{pmatrix} 0.9772 & 0.0228 \\ 0.0446 & 0.9554 \end{pmatrix}$

ple which retains only methylated DNA has a lower mean intensity than the sample which retains only unmethylated DNA, and vice versa is true for the means of methylated probes ($\hat{\mu}_1$). As in the Lippman et al. (2004) data, the correlation between two samples collected from the same individual is near 0.8.

6.3 Comparison of Chromosome 4 Results

DNA methylation in wild-type Columbia *Arabidopsis* is investigated in both studies presented above. Since a whole genome tiling array is employed in the Zhang

Table 6.2
Parameter estimates for the bivariate normal observation probability distribution (4.15), obtained from the modified BW algorithm for the annotated HMM. Parameter estimates in the left column are for unmethylated probes, while the estimates on the right column are for methylated probes. Results are given for each of the five *Arabidopsis* chromosomes.

Chr.	$\hat{\mu}_0$	$\hat{\sigma}_{01}, \hat{\sigma}_{02}$	$\hat{\rho}_0$	$\hat{\mu}_1$	$\hat{\sigma}_{11}, \hat{\sigma}_{12}$	$\hat{\rho}_0$
1	$\begin{pmatrix} 3.8725 \\ 4.4338 \end{pmatrix}$	1.5740, 1.6314	0.8561	$\begin{pmatrix} 5.5422 \\ 4.4821 \end{pmatrix}$	1.6724, 1.7654	0.8200
2	$\begin{pmatrix} 3.9442 \\ 4.4710 \end{pmatrix}$	1.6313, 1.6883	0.8670	$\begin{pmatrix} 5.5254 \\ 4.3507 \end{pmatrix}$	1.6658, 1.7329	0.8073
3	$\begin{pmatrix} 3.9369 \\ 4.4746 \end{pmatrix}$	1.5995, 1.6453	0.8589	$\begin{pmatrix} 5.5214 \\ 4.3875 \end{pmatrix}$	1.6582, 1.7520	0.8148
4	$\begin{pmatrix} 3.8947 \\ 4.4307 \end{pmatrix}$	1.5860, 1.6375	0.8585	$\begin{pmatrix} 5.5794 \\ 4.4600 \end{pmatrix}$	1.6742, 1.7761	0.8216
5	$\begin{pmatrix} 3.8661 \\ 4.4204 \end{pmatrix}$	1.5693, 1.6296	0.8577	$\begin{pmatrix} 5.4997 \\ 4.4930 \end{pmatrix}$	1.6669, 1.7990	0.8292

et al. (2006) study, the probes which correspond to the chromosome 4 heterochromatic knob can be compared to results from the Lippman et al. (2004) study. First, a few design differences between the whole genome array and the custom-designed cDNA array are noteworthy. On the Affymetrix® whole genome array, all probes are 25 base pairs long, with an average gap of 10 bases between probes. Even though the Affymetrix® tiling array does not include probes in repetitive DNA regions, which are known to be present in the heterochromatic knob, a total of 14,714 probes cover the region. This corresponds to 367,850 nucleotide bases, which is 50% of the bases

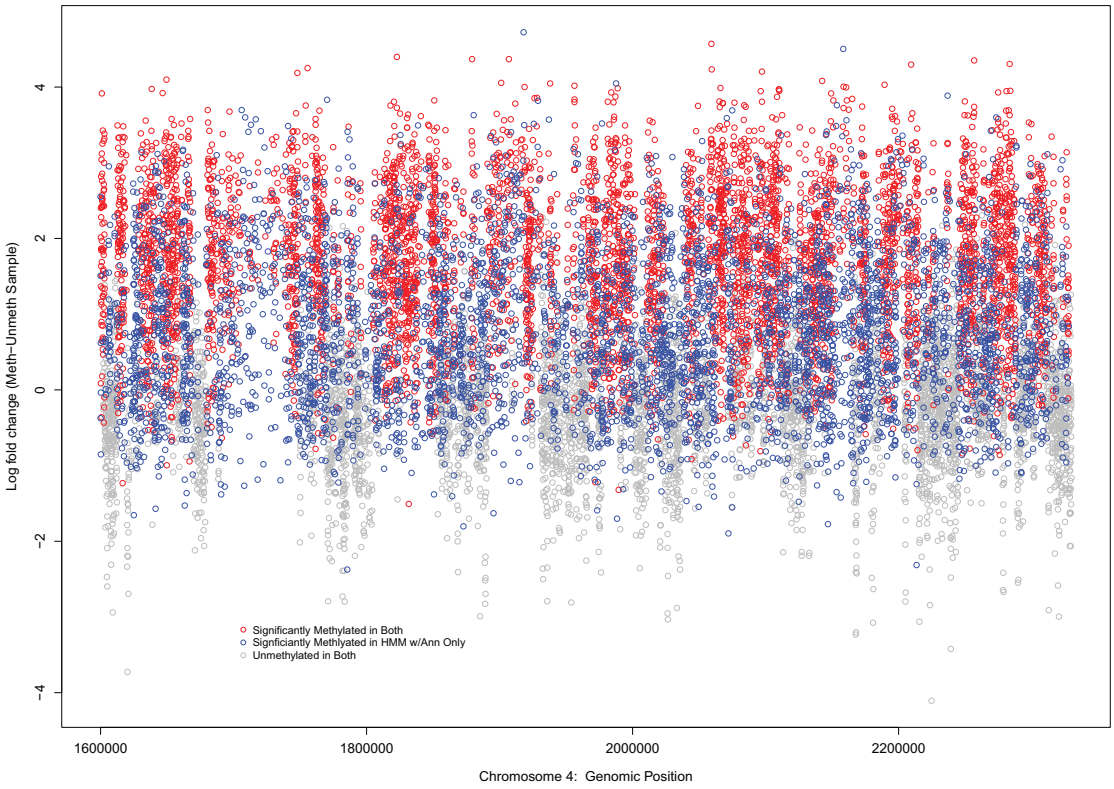

Figure 6.6. Chromosome 4 heterochromatic knob results from Affymetrix® whole genome tiling arrays. Log fold change (methylated DNA sample - unmethylated DNA sample) is plotted against the genomic position in the knob region. Red points are significantly methylated points using both the HMM with genomic annotation and Tilemap, blue points are probes that were only identified as methylated with the annotated HMM, and grey points are unmethylated probes using both methods. Note that no points were identified as methylated with Tilemap that were not also found in the HMM with annotation.

in the whole knob region (that is 730,000 nucleotide bases in length). By comparison, the cDNA tiling array custom-designed by Lippman et al. (2004) covers the entire knob region, but probes are longer (an average of 995 bases long) and vary in length. These design differences make it difficult to make a direct comparison between the

two studies, but it is worthwhile to investigate the results from the Zhang et al. (2006) study in that region.

Figure 6.6 shows DNA methylation status results from the annotated HMM and Tilemap in the chromosome 4 heterochromatic knob region using Affymetrix® whole genome arrays. The log fold change between the methylated and unmethylated DNA sample is given on the y-axis. It is clear that while both methods (red points) identify many methylated probes in the region, the HMM with annotation (blue points) identifies almost twice as many (70%) methylated probes in the region as does Tilemap (37.78%). While it is clear from these results that there is heavy methylation in this region, there is a noticable presence of gaps in coverage (e.g., around genomic position 1,700,000). This is due to the lack of repetitive DNA coverage on the Affymetrix® whole genome tiling array. Although the results from the whole genome array appear to be consistent with the heavy pattern of DNA methylation shown in Lippman et al. (2004), the Affymetrix® tiling arrays do not truly cover the whole knob region and DNA methylation cannot be investigated at those locations where gaps in coverage exist.

6.4 Summary

In this chapter two real DNA methylation tiling array data sets for *Arabidopsis thaliana* are analyzed. In the first study (Lippman et al. (2004)), in which a custom-designed tiling array of the heterochromatic knob of chromosome 4 is utilized, the HMM which incorporates genomic annotation is compared to an ANOVA previously used to analyze the data. Both methods detect heavy methylation in the knob region, which is expected. However, the HMM with annotation identifies some methylated probes at the right end of the heterochromatic knob that the ANOVA model does not identify. In the second study (Zhang et al. (2006)), in which an Affymetrix® whole genome tiling array is employed, the HMM with genomic annotation is compared to the HMM used in Tilemap (Ji and Wong, 2005), which Zhang et al. (2006) uses

to detect methylated probes. The HMM which uses genomic annotation identified over twice as many methylated probes (mostly in gene regions) that Tilemap did not. Finally, although the transition probabilities are similar between genes and intergenic regions in both studies, there are some slight differences, and in the Zhang et al. (2006) study, the annotated HMM identified several significantly methylated probes that the unannotated model did not. Most importantly, use of the annotated HMM allows for the investigation of different patterns in genes and intergenic regions, which has not been explored in other methods. Since the annotated HMM performs well, even in the case when transition probabilities are the same in genes and intergenic regions, it is beneficial to employ this method in DNA methylation tiling array studies since it allows for the possibility of differences by genomic elements and enables the investigation of patterns in those regions.

7. SUMMARY AND FUTURE WORK

7.1 Summary

Understanding the factors that affect phenotypic variation is a complex task that is central to the fields of genetics and epigenetics. The study of genetics reveals how changes or differences in the DNA sequence can lead to phenotypic differences, while the field of epigenetics addresses heritable changes to phenotypes or gene expression that are not due to a change in the DNA sequence alone. In the 1990s and early 2000s, advances in technology made it possible to move from small-scale genetic studies to the investigation of whole genomes. Genome sequencing projects for many organisms started to flourish with the goal of determining all of an organism's nucleotide base pairs, along with identifying genes and other genomic elements. Online genomic databases were created to store this genomic annotation information in a publicly available manner (Stein, 2001). Along with the sequence and annotation information, came the development of microarray technology (Schena et al., 1995), initially used to monitor the expression levels of thousands of genes in a single experiment. Advances in microarray technology made possible the study of the whole genome (not just gene regions), enabling the large-scale investigation of epigenetic modifications that can occur anywhere in genome. As technologies continue to advance, new insights into the fields of genomics and epigenomics will continue to challenge researchers to best utilize available information to gain a better understanding of biological processes.

In this research, incorporation of genomic annotation information into the statistical analysis of data obtained from a specific microarray technology called a tiling array is explored. Tiling arrays offer unbiased coverage of entire genomic regions (often whole genomes) through the sequential placement of probes from one end of the region to the other. Due to their dense coverage, tiling arrays have been used in a

variety of applications, such as transcription mapping, identification of transcription factor binding sites, DNA methylation profiling, and investigation of histone modifications (Mockler and Ecker, 2005). Although selection of probes without regard to genomic annotation is essential to their broad applicability, knowing which probes correspond to which genomic elements (i.e, genes, introns, exons) is useful information that is often used after a statistical analysis to visualize results in terms of genomic annotation. In this work, genomic annotation information is incorporated into the statistical analysis of two different applications of tiling array data: differential expression analysis and DNA methylation profiling. The process of connecting tiling array probes to genomic annotation is described in Chapter 2, and results from the annotation of two different tiling array designs for the model plant *Arabidopsis thaliana* are given in detail. These two tiling arrays are the basis for real data applications explored in this research.

In particular, differential expression analysis using an *Arabidopsis* Affymetrix® whole genome tiling array is described in Chapter 3. Differential expression analysis involves the study of expression differences between conditions of interest (e.g, treatment vs. control) for all known genes. While statistical issues for studying differential expression using gene expression microarray technology have been thoroughly investigated (e.g., Kerr et al. (2000); Wolfinger et al. (2001); Irizarry et al. (2003); Smyth (2004)), relatively few studies employ tiling arrays for differential expression analysis. Although they are not specifically designed to optimize the study of gene expression, the dense coverage provided by tiling arrays offers the potential for coverage of recently discovered genes, as well as provides more coverage per gene. However, since gene expression is expected to occur in exon regions of genes, many of the tiling array probes are not biologically relevant. Without genomic annotation information, statistical testing is limited to individual probe-level testing since it is unknown which probes correspond to genes. However, by using genomic annotation information to filter out probes in introns and intergenic regions, a set of biologically relevant probes is obtained that corresponds to the same format as gene expression microarray data

with multiple probes located in the exon regions of genes. Performing this initial bioinformatic step allows the use of statistical methods previously developed for gene expression microarrays. An ANOVA model (Wolfinger et al., 2001; Chu et al., 2002) is applied to gene expression data obtained by hybridizing mRNA samples collected from the same individuals to both gene expression and tiling microarrays. Results from this study indicate that (for genes represented on both arrays) although there is some overlap among differentially expressed genes, some sizable differences are seen in gene expression level measurements. While more work needs to be done to gain a better understanding of how the design differences between gene expression and tiling arrays affects the measurement of gene expression, this application demonstrates how genomic annotation (i.e, knowing the locations of which probes are in exon regions of genes) can lead to a biologically relevant statistical analysis.

In Chapter 4, a method is proposed for incorporating genomic annotation information into DNA methylation profiling studies. DNA methylation is an epigenetic mechanism that occurs when a methyl chemical group attaches to a cytosine base on the DNA sequence. DNA methylation is important in the regulation of gene expression and has been shown to be associated with many types of cancer (Jaenisch and Bird, 2003; Jones and Baylin, 2007). To better understand this epigenetic mechanism, it is imperative to identify locations of DNA methylation in a genome for a variety of organisms and cell types. Genome-wide profiling of DNA methylation is feasible with the use of tiling arrays. Early studies employ ANOVA models (Lippman et al., 2004; Vaughn et al., 2007) or sliding window approaches (Cawley et al., 2004; Keles et al., 2006) to identify which probes on the tiling array are significantly methylated. However, both of these approaches require the testing of thousands (if not millions) of statistical tests that are not likely to be independent. More recently, hidden Markov models (HMMs) (Ji and Wong, 2005; Du et al., 2006; Humburg et al., 2008; Yoo, 2008) have successfully been employed in DNA methylation tiling array experiments to estimate the DNA methylation status of all probes while incorporating dependency between neighboring probes. However, many of these models employ *ad hoc* param-

eter estimation and do not make use of genomic annotation information. Previous studies have offered insight into DNA methylation patterns of many organisms and, in some cases, seem to indicate that patterns may differ by genomic element (e.g., gene, transposon)(Zhang et al., 2006; Suzuki and Bird, 2008). These findings motivate the idea of incorporating genome annotation information into a HMM framework to improve prediction of DNA methylation. A method is proposed, in which transition probabilities are allowed to vary between gene and intergenic regions. The forward-backward (FB) algorithm (employed for state estimation)(Baum et al., 1970; Baum, 1972) and the Baum-Welch algorithm (an EM-algorithm for HMMs used to obtain maximum likelihood estimates of model parameters) (Baum et al., 1970) are both modified to accommodate these changes in transition probabilities according to genomic element.

Simulation studies are employed in Chapter 5 to investigate the use of genomic annotation in DNA methylation profiling studies with tiling arrays. Data are simulated for two different scenarios. In the first case, transition probabilities are assumed to be different for gene and intergenic regions, and, in the second case, transition probabilities are assumed to be the same across the whole region. Different parameter settings for the observation probability distribution are investigated to gain an understanding of model performance across a wide range of parameter values. In the first simulation study, HMM model parameters are assumed to be known and both the standard (unannotated) and modified (annotated) forward-backward algorithms are applied to estimate hidden states. This study reveals that if there truly is a difference in transition probabilities between gene and intergenic regions, then incorporating genomic annotation information results in a greater proportion of correctly predicted states than if transition probabilities are assumed to be constant across the region. While the annotated model outperforms the unannotated model across all parameter settings, it is noteworthy that the difference in model performance is most drastic when the data are more noisy and/or for smaller mean differences between intensity levels of the untreated and treated samples. When there truly are no differences in tran-

sition probabilities between genes and intergenic regions, performance of annotated and unannotated model is identical. In the second simulation study, HMM model parameters are estimated from the data using the standard (unannotated) and modified (annotated) versions of the Baum-Welch algorithm. These models are compared to results from paired t-tests and FB algorithm implementation with *ad hoc* parameter estimates. Both of the latter two models perform poorly across most parameter settings, while estimating hidden states via FB with BW parameter estimates shows a marked improvement over these two methods. This indicates the importance of both modeling dependency between neighboring probes and utilizing a formal estimation procedure for HMM model parameters. Additionally, when there truly is a difference in transition probabilities between genes and intergenic regions, incorporating genomic annotation information into the estimation algorithms, yields model performance that is similar to that if the true parameters were known. Again, utilizing the annotated model when there truly are no differences in transition probabilities does not affect model performance.

In Chapter 6, the proposed method for incorporating genomic annotation into the HMM framework for DNA methylation tiling array studies is applied to two real data sets. The first data set (Lippman et al., 2004) investigates DNA methylation in the region surrounding the heterochromatic knob on *Arabidopsis* chromosome 4. An ANOVA model based on the original analysis is employed to test for significantly methylated probes across a region of 1407 probes and results are compared to the HMM which utilizes genomic annotation information. Results are similar between the two methods, with the annotated HMM detecting a few more methylated probes at the right end of the heterochromatic knob. The second data set (Zhang et al., 2006) explores DNA methylation across the whole *Arabidopsis* genome by using Affymetrix® whole genome tiling arrays. A HMM approach developed by Ji and Wong (2005) is employed in this study to identify methylated probes and these results are compared to those from the annotated HMM. While many of the same probes are identified as methylated using both approaches, the annotated HMM detects about twice as

many methylated probes. In both studies, the annotated HMM is also compared to the unannotated HMM. The differences in transition probabilities are small in magnitude, leading to a few differences in DNA methylation status prediction in the Zhang et al. (2006) study. However, it is only through the use of the annotated HMM that differences in transition probabilities can be explored and since this model always allows for the potential of such differences (without affecting model performance if they do not exist), it is an important contribution to DNA methylation profiling with tiling arrays.

In summary, genomic annotation information can be used in the statistical analysis of tiling array data. In this work, the use of genomic annotation information is applied to differential expression analysis and DNA methylation profiling. For differential expression analysis, genomic annotation is used to perform a filtering step to identify biologically relevant probes and obtain data in the same form as that of gene expression microarrays. For DNA methylation profiling studies, genomic annotation is incorporated into a HMM framework by allowing different transition probabilities for genes and intergenic regions. Modified versions of the standard HMM estimation algorithms (FB and BW) are developed to incorporate these differences in transition probabilities with the goal of improving prediction of DNA methylation status.

7.2 Future Directions

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 transposable elements or promoter regions of genes) 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, instead of assuming a constant set of transition probabilities across all genes, these transitions could be allowed to vary by gene. While the methods for DNA methylation profiling studies described in this work ad-

dress the identification of locations of DNA methylation in one sample type (e.g., wild-type Columbia *Arabidopsis*), it is often of interest to compare DNA methylation status across sample types (e.g., disease vs. healthy individuals). This presents another opportunity to investigate the use of genomic annotation since perhaps different sample types could have different sets of transition probabilities for different genomic elements.

In most DNA methylation profiling studies (including the simulations described in this work), a common underlying DNA methylation status for each probe is assumed across all biological replicates. Variation between individuals is assumed to be in the observed data and not in the underlying hidden states. While this assumption makes it reasonable to take the average of the observed data for each biological replicate, the assumption of a common underlying state may not be realistic for real data. Estimating methylation status of all probes for each biological replicate and determining the overall methylation status across replicates may be of interest. In this case, taking the average of the biological replicates may not be ideal since the observed data of different individuals may come from different underlying distributions. Employing an estimation technique which allows the direct use of each individual's observed data to be combined in a way that accounts for the potential variation in underlying hidden states is desirable. Such techniques have been described in the HMM literature as methods which employ the use of multiple observation sequences (Rabiner, 1989). It is of interest to investigate this method of estimation on simulated and real data in which hidden states are not assumed to be the same across all replicates.

This work extends the standard forward-backward (FB) and Baum-Welch (BW) algorithms for hidden state and parameter estimation within the hidden Markov model (HMM) framework to identify the DNA methylation status for all tiling array probes. The modifications to the FB and BW algorithms allow the incorporation of genomic annotation via estimation of separate transition probabilities for genes and intergenic regions. Although statistical properties of the standard FB and BW algorithms are well-understood, convergence properties of the modified algorithms

need to be thoroughly explored. One limitation of the standard BW algorithm is that it may converge to a local maximum of the likelihood function. Bayesian methods offer an alternative to the FB and BW algorithms for hidden state and parameter estimation. Within the HMM framework, priors could be assigned to the model parameters $\lambda = (A, B, \pi)$ and the hidden state sequence. Mo and Liang (2010) present one option of a Bayesian analysis within the HMM context where the hidden state vector is modeled by a one-dimensional Ising model, which is a Markov random field model. Although this method does not use genomic annotation, such information can potentially be incorporated into Bayesian methods via the prior specifications. Coupling genomic annotation information with Bayesian approaches for modeling DNA methylation profiling data offers a potentially powerful alternative to the HMM framework presented in this work and merits further investigation.

Finally, although the methods in this work are designed for tiling arrays, a newer type of technology, referred to as next generation sequencing, has become a popular way to study many different types of biological phenomena, including gene expression and DNA methylation. Data generated from next generation sequencing technologies is in the form of counts rather than continuous intensity measurements obtained from microarrays. Thus, one place to begin an extension of the HMM for DNA methylation profiling studies is in the specification of the observation probability distribution. Investigating how the methods in this work can be extended to next generation sequencing studies will be an important endeavor in advancing this work.

LIST OF REFERENCES

LIST OF REFERENCES

Affymetrix® Data Sheet (2004). GeneChip Arabidopsis ATH1 genome array. http://www.affymetrix.com/support/technical/datasheets/arab_datasheet.pdf.

Affymetrix® Image Library (2009). GeneChip single feature and GeneChip hybridization images. www.affymetrix.com.

Affymetrix® Package Insert (2006). GeneChip Arabidopsis tiling 1.0R array. https://www.affymetrix.com/support/downloads/package_inserts/tiling_arabidopsis_insert.pdf.

Affymetrix® Technical Note (2007). Array design for the GeneChip human genome U133 set. www.affymetrix.com/support/technical/technotes/hgu133_design_technote.pdf.

AHEAD (2008). Moving AHEAD with an international human epigenome project. *Nature 454*, 711–715.

Allfrey, V. G., R. Faulkner, and A. E. Mirsky (1964). Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. *Proceedings of the National Academy of Science USA 51*, 786–794.

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.

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler (2008). GenBank. *Nucleic Acids Research 36*, D25–D30.

Berger, S. L. (2002). Histone modifications in transcriptional regulation. *Current Opinion in Genetics and Development 12*, 142–148.

Berger, S. L., T. Kouzarides, R. Shiekhattar, and A. Shilatifard (2009). An operational definition of epigenetics. *Genes and Development 23*, 781–783.

Bertone, P., V. Stolc, T. E. Royce, J. S. Rozowsky, A. E. Urban, X. Zhu, J. L. Rinn, W. Tongprasit, M. Samanta, S. Weissman, M. Gerstein, and M. Snyder (2004). Global identification of human transcribed sequences with genome tiling arrays. *Science 306*, 2242–2246.

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

Bird, A. (2007). Perceptions of epigenetics. *Nature 447*, 396–398.

Black, M. A. and R. W. Doerge (2002). Calculation of the minimum number of replicate spots required for detection of significant gene expression fold change in microarray experiments. *Bioinformatics 18*, 1609–1616.

Bolstad, B., R. Irizarry, M. Astrand, and T. Speed (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics 19*, 185–193.

Brownwell, J., J. Zhou, T. Ranalli, R. Kobayashi, D. Edmondson, S. Roth, and C. D. Allis (1996). *Tetrahymena* histone acetyltransferase A: a homolog to yeast Gcn5p linking histone acetylation to gene activation. *Cell 84*, 843–851.

Buck, M. J. and J. D. Lieb (2004). ChIP-chip: consideration for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics 83*, 349–360.

Buck, M. J., A. B. Nobel, and J. D. Leib (2005). ChiPOTle: a user friendly tool for the analysis of ChIP-chip data. *Genome Biology 6*, R97.

Cappe, O., E. Moulines, and T. Ryden (2005). *Inference in hidden Markov models*. Springer.

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.

Cedar, H. and Y. Bergman (2009). Linking DNA methylation and histone modification: patterns and paradigms. *Nature Reviews Genetics 10*, 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*, 351–360.

Chu, T.-M., B. Weir, and R. Wolfinger (2002). A systematic statistical linear modeling approach to oligonucleotide array experiments. *Mathematical Biosciences 176*, 35–51.

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

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 39*, 1–38.

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.

Duggan, D. J., M. Bittner, Y. Chen, P. Meltzer, and J. M. Trent (1999). Expression profiling using cDNA microarray. *Nature Genetics 21*, 10–14.

Durbin, R., S. Eddy, A. Krogh, and G. Mitchison (1998). *Biological Sequence Analysis*. Cambridge University Press.

Estecio, M. R. and J.-P. J. Issa (2009). Tackling the methylome: recent methodological advances in genome-wide methylation profiling. *Genome Medicine 1*, 106.

Feinberg, A. P. and B. Tycko (2004). The history of cancer epigenetics. *Nature Reviews Cancer 4*, 143–153.

Feinberg, A. P. and B. Vogelstein (1983). Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature 301*, 89–92.

Finnegan, E. J. (2010). DNA methylation: a dynamic regulator of genome organization and gene expression in plants. In *Plant Developmental Biology - Biotechnological Perspectives*. Springer Berlin Heidelberg.

Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, K. McKenney, G. Sutton, W. Fitzhugh, C. Fields, J. D. Gocayne, J. Scott, R. Shirley, L. L. Liu, A. Glodek, J. M. Kelley, J. F. Weidman, C. A. Phillips, T. Spriggs, E. Hedblom, M. D. Cotton, T. R. Utterback, M. C. Hanna, D. T. Nguyen, D. M. Saudek, R. C. Brandon, L. D. Fine, J. L. Fritchman, J. L. Fuhrmann, N. S. M. Geoghagen, C. L. Ghehm, L. A. McDonald, K. V. Small, C. M. Fraser, H. O. Smith, and J. C. Venter (1995). Whole-genome random sequencing and assembly of *haemophilus influenza* Rd. *Science 269*, 496–512.

Forney, G. D. (1973). The Viterbi algorithm. *Proceedings of the IEEE 61*, 268–278.

Gehring, M. and S. Henikoff (2007). DNA methylation dynamics in plant genomes. *Biochimica et Biophysica Acta 1769*, 276–286.

Gelvin, S. B. (2003). Agrobacterium and plant transformation: the biology behind the "gene-jockeying" tool. *Microbiology and Molecular Biology Reviews 67*, 16–37.

Gentleman, R. C., V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology 5*, R80.

Ghosh, S., H. A. Hirsch, E. A. Sekinger, P. Kapranov, K. Struhl, and T. R. Gingeras (2007). Differential analysis for high density tiling microarray data. *BMC Bioinformatics 8*, 359.

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

Gottardo, R., W. Li, W. E. Johnson, and X. S. Liu (2008). A flexible and powerful Bayesian hierarchical model for ChIP-chip experiments. *Biometrics 64*, 468–478.

Grewal, S. I. S. and S. Jia (2007). Heterochromatin revisited. *Nature Reviews Genetics 8*, 35–46.

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

Hazen, S. P., F. Naef, T. Quisel, J. M. Gendron, H. Chen, J. R. Ecker, J. O. Borevitz, and S. A. Kay (2009). Exploring the transcriptional landscape of plant circadian rhythms using genome tiling arrays. *Genome Biology 10*, R17.

Herman, J. G. and S. B. Baylin (2003). Gene silencing in cancer in association with promoter hypermethylation. *The New England Journal of Medicine 349*, 2042–2054.

Holliday, R. (2006). Epigenetics: a historical overview. *Epigenetics 1*, 76–80.

Holliday, R. and J. E. Pugh (1975). DNA modification mechanisms and gene activity during development. *Science 187*, 226–232.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics 6*, 65–70.

Huber, W., J. Toedling, and L. M. Steinmetz (2006). Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics 22*, 1963–1970.

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

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature 409*, 860–921.

Irizarry, R. A., B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics 4*, 249–264.

Jaenisch, R. and A. Bird (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genetics 33*, 245–254.

Jelinek, F. (1997). *Statistical methods for speech recognition*. MIT Press.

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.

Jones, P. A. and R. Martienssen (2006). A blueprint for a Human Epigenome Project: the AACR human epigenome workshop. *Cancer Research 65*, 11241–11246.

Juang, B. and L. Rabiner (1990). A segmental k-means algorithm for estimating parameters of hidden Markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing 38*, 1639–1641.

Kampa, D., J. Cheng, P. Kapranov, M. Yamanaka, S. Brubaker, S. Cawley, J. Drenkow, A. Piccolboni, S. Bekiranov, G. Helt, H. Tammana, and T. R. Gingeras (2004). Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Research 14*, 331–342.

Kapranov, P., S. E. Cawley, J. Drenkow, S. Bekiranov, R. L. Strausberg, S. P. Fodor, and T. R. Gingeras (2002). Large-scale transcriptional activity in chromosomes 21 and 22. *Science 296*, 916–919.

Keles, S. (2007). Mixture modeling for genome-wide localization of transcription factors. *Biometrics 63*, 10–21.

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.

Kerr, M. K., M. Martin, and G. A. Churchill (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology 7*, 819–837.

Keshet, I., Y. Schlesinger, S. Farkash, E. Rand, M. Hecht, E. Segal, E. Pikarski, R. A. Young, A. Niveleau, H. Cedar, and I. Simon (2006). Evidence for an instructive mechanism of de novo methylation in cancer cells. *Nature Genetics 38*, 149–153.

Kim, J. K., M. Samaranayake, and S. Pradhan (2009). Epigenetic mechanisms in mammals. *Cellular and Molecular Life Sciences 66*, 596–612.

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

Kouzarides, T. and S. L. Berger (2007). Chromatin modifications and their mechanism of action. In *Epigenetics*. Cold Spring Harbor Laboratory Press.

Kuo, M.-H. and C. D. Allis (1998). Roles of histone acetyltransferases and deacetylases in gene regulation. *BioEssays 20*, 615–626.

Lashkari, D. A., J. L. DeRisi, J. H. McCusker, A. F. Namath, C. Gentile, S. Y. Hwang, P. O. Brown, and R. W. Davis (1997). Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proceedings of the National Academy of Sciences of the United States of America 94*, 13057–13062.

Law, J. A. and S. E. Jacobsen (2010). Establishing, maintaining, and modifying DNA methylation patterns in plants and animals. *Nature Reviews Genetics 11*, 204–220.

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.

Liolios, K., i. M. Chen, K. Mavromatis, N. Tavernarakis, P. Hugenholtz, V. M. Markowitz, and N. C. Kyrpides (2010). The Genomes OnLine Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research 38*, D346–D354.

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.

Lipshutz, R. J., S. A. Fodor, T. R. Gingeras, and D. J. Lockhardt (1999). High density synthetic oligonucleotide arrays. *Nature Genetics 21*, 20–24.

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.

Margueron, R. and D. Reinberg (2010). Chromatin structure and inheritance of epigenetic information. *Nature Reviews Genetics 11*, 285–296.

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.

Meinke, D. W., J. M. Cherry, C. Dean, S. D. Rounsley, and M. Koornneef (1998). Arabidopsis thaliana: a model plant for genome analysis. *Science 282*, 662–682.

Mo, Q. and F. Liang (2010). A hidden Ising model for ChIP-chip data analysis. *Bioinformatics 26*, 777–783.

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

Munch, K., P. P. Gardner, P. Arctander, and A. Krogh (2006). A hidden Markov model approach for determining expression from genomic tiling micro arrays. *BMC Bioinformatics 7*, 239.

Naouar, N., K. Vandepoele, T. Lammens, T. Casneuf, G. Zeller, P. van Hummelen, D. Weigel, G. Ratsch, D. Inze, M. Kuiper, L. D. Veylder, and M. Vuylsteke (2009). Quantitative RNA expression analysis with Affymetrix tiling 1.0R arrays identifies new E2F target genes. *The Plant Journal 57*, 184–194.

National Center for Biotechnology Information (2010). Genbank. http://www.ncbi.nlm.nih.gov/genbank/index.html.

Olbricht, G. R., N. Sardesai, S. B. Gelvin, B. A. Craig, and R. W. Doerge (2009). Statistical methods for Affymetrix tiling array data. In *The Proceedings of the Kansas State University Conference on Applied Statistics in Agriculture*.

Qi, Y., A. Rolfe, K. D. MacIsaac, G. K. Gerber, D. Pokholok, J. Zeitlinger, T. Danford, R. D. Dowell, E. Fraenkel, T. S. Jaakkola, R. A. Young, and D. K. Gifford (2006). High-resolution computational models of genome binding events. *Nature Biotechnology 24*, 963–970.

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

R Core Development Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://R-project.org.

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.

Rau, A. (2010). *Reverse engineering gene networks using genomic time-course data*. Ph.D. dissertation, Purdue University, West Lafayette, IN USA.

Riggs, A. D. (1975). X inactivation, differentiation, and DNA methylation. *Cytogenetics and cell genetics 14*, 9–25.

Robertson, K. D. (2005). DNA methylation and human disease. *Nature Reviews Genetics 6*, 597–610.

Russo, V., R. Martienssen, and A. Riggs (Eds.) (1996). *Epigenetic mechanisms of gene regulation*. Cold Spring Harbor Laboratory Press.

Schadt, E. E., S. W. Edwards, D. GuhaThakurta, D. Holder, L. Ying, V. Svetnik, A. Leonardson, K. W. Hart, A. Russell, G. Li, G. Cavet, J. Castle, P. McDonagh, Z. Kan, R. Chen, A. Kasarskis, M. Margarint, R. M. Caceres, J. M. Johnson, C. D. Armour, P. W. Garrett-Engele, N. F. Tsinoremas, and D. D. Shoemaker (2004). A comprehensive transcript index of the human genome generated using microarrays and computational approaches. *Genome Biology 5*, R73.

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.

Schumacher, A., P. Kapranov, Z. Kaminsky, J. Flanagan, A. Assadzadeh, P. Yau, C. Virtanen, N. Winegarden, J. Cheng, T. Gingeras, and A. Petronis (2006). Microarray-based DNA methylation profiling: technology and applications. *Nucleic Acids Research 34*, 528–542.

Shames, D. S., J. D. Minna, and A. F. Gazdar (2007). DNA methylation in health, desease, and cancer. *Current Molecular Medicine 7*, 85–102.

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

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology 3*, Article 3.

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.

The Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature 408*, 796–815.

The Arabidopsis Information Resource (TAIR) (2008). http://www.arabidopsis.org/.

Vaillant, I. and J. Paszkowski (2007). Role of histone and DNA methylation in gene regulation. *Current Opinion in Plant Biology 10*, 528–533.

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.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. G. Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. D. Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R.-R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Y. Wang, A. Wang, X. Wang, J. Wang, M.-H. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. C. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y.-H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guig, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y.-H. Chiang, M. Coyne, C. Dahlke, A. D. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu (2001). The sequence of the human genome. *Science 291*, 1304–1351.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Transactions on Information Theory 13*, 260–269.

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

Watson, J. D. and F. H. C. Crick (1953). A structure for deoxyribose nucleic acid. *Nature 171*, 737–738.

Weber, M., J. J. Davies, D. Wittig, E. J. Oakeley, M. Haase, W. L. Lam, and D. Schubeler (2005). Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nature Genetics 37*, 853–862.

Wolfinger, R. D., G. Gibson, E. D. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari, and R. S. Paules (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology 8*, 625–637.

Wu, M., F. Liang, and Y. Tian (2009). Bayesian modeling of ChIP-chip data using latent variables. *BMC Bioinformatics 10*, 352.

Yamada, K., J. Lim, J. M. Dale, H. Chen, P. Shinn, C. J. Palm, A. M. Southwick, H. C. Wu, C. Kim, M. Nguyen, P. Pham, R. Cheuk, G. Karlin-Newmann, S. X. Liu, B. Lam, H. Sakano, T. Wu, G. Yu, M. Miranda, H. L. Quach, M. Tripp, C. H. Chang, J. M. Lee, M. Toriumi, M. M. H. Chan, C. C. Tang, C. S. Onodera, J. M. Deng, K. Akiyama, Y. Ansari, T. Arakawa, J. Banh, F. Banno, L. Bowser, S. Brooks, P. Carninci, Q. Chao, N. Choy, A. E. andAndrew D. Goldsmith, M. Gurjal, N. F. Hansen, Y. Hayashizaki, C. Johnson-Hopson, V. W. Hsuan, K. Iida, M. Karnes, S. Khan, E. K. andJunko Ishida andPaul X. Jiang, T. J. andJun Kawai, A. Kamiya, C. Meyers, M. Nakajima, M. Narusaka, M. Seki, T. Sakurai, M. Satou, R. Tamse, M. Vaysberg, E. K. Wallender, C. Wong, Y. Yamamura, S. Yuan, K. Shinozaki, R. W. Davis, A. Theologis, and J. R. Ecker (2003). Empirical analysis of transcriptional activity in the Arabidopsis genome. *Science 302*, 842–846.

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

Yoo, S.-Y. and R. Doerge (2009). *Epigenomics*, Chapter Integrating epigenomic results, pp. 37–53. Springer.

Zeller, G., S. R. Henz, C. K. Widmer, T. Sachsenberg, G. Ratsch, D. Weigel, and S. Laubinger (2009). Stress-induced changes in the Arabidopsis thaliana transcriptome analyzed using whole-genome tiling arrays. *The Plant Journal 58*, 1068–1082.

Zhang, X., O. Clarenz, S. Cokus, Y. V. Bernatavichute, M. Pellegrini, J. Goodrich, and S. E. Jacobsen (2007). Whole-genome analysis of histone H3 lysine 27 trimethylation in Arabidopsis. *PLoS Biology 5*, e129.

Zhang, X., S. Shiu, A. Cai, and J. O. Borevitz (2008). Global analysis of genetic, epigenetic, and transcriptional polymorphisms in Arabidopsis thaliana using whole genome tiling arrays. *PLoS Genetics 4*, e1000032.

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.

VITA

VITA

Gayla R. (Hobbs) Olbricht was born in West Plains, Missouri, USA on September 10, 1979. She received an A.A. in General Studies in 1999 followed by a B.S. in Mathematics in 2001 from Missouri State University. In 2004, Gayla received a M.S. in Applied Statistics from Purdue University and has been a graduate student in the Purdue Statistics Department since 2002.