Spring 5-15-2017

Modeling Complex Patterns of Differential DNA Methylation That Associate with Expression Change

Christopher E. Schlosberg
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Computational and Systems Biology

Dissertation Examination Committee:
Gary Stormo, Chair
John Edwards
Tao Ju
Christopher Maher
Eugene Oltz

Modeling Complex Patterns of Differential DNA Methylation That Associate with Expression Change

By
Christopher Schlosberg

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2017
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would first like to thank the support of the Washington University in St. Louis community in the Division of Biology and Biomedical Sciences and the Computational and Systems Biology Program for the chance to complete my PhD research. I am grateful for the financial support and guidance through advisors Drs. Barak Cohen, Michael Brent, Ting Wang, and Gautam Dantas and counselors Jeanne Silvestrini and Melanie Relich. GATP supported my research for three years and provided resources for me to attend the 22nd Annual International Conference on Intelligent Systems for Molecular Biology.

I would not have completed this dissertation without the practical wisdom of my advisor Dr. John Edwards. John has steadily promoted not only my academic development, but also my professional development. He supported me when I took a larger role in our science policy graduate student group ProSPER and when I accepted a six month leave of absence to join Monsanto Company. These opportunities have shaped me as a scientist and as a person, and I am grateful his door has always been open to me as a student and as a friend. I am thankful for the guidance and friendship of Dr. Nathan VanderKraats, who not only was instrumental in the early development of my thesis, but consistently encouraged my abilities, curiosity, and professional development. I have received thoughtful guidance from Drs. Kilian Weinberger and Tao Ju in machine learning applications and algorithm design. I am thankful for all the critical discussions that Drs. Gary Stormo, Eugene Oltz, and Christopher Maher have provided to enhance my thesis.

I am thankful for the scientific discussion and debate with other present and past lab members: James McDonald, Jerry Fong, Lisa Rois, Dr. Jeff Hiken, Dr. Keith Decker, Dr. Manoj Singh,

Margaret Akinhanmi, Alexis Fennoy, Geoffrey Cheng, Kyle Jung, Tolison Fowler, Pooja Tripathy. I appreciate the comradery and scientific input from other members of the Center for Pharamacogenomics: Dr. Cristina De Guzman Strong, Dr. Gerald Dorn, Dr. Scot J. Matkovich, Dr. Li Jia, Dr. Tami Bowman, Dr. Inez Oh, Dr. Ashley Quiggle, Zane Goodwin, Mary Mathyer. I am grateful for the friendships that I have made in DBBS over the years: Drs. Kilannin Krysiak, Michael Stevens, Ruteja Barve, Nic Ho, and Vasavi Sundaram.

I thank my parents Nancy Lau and Neal Schlosberg, my grandparents Elizabeth and Edward Lau, Barbara and Alexander Schlosberg, and my brother Phillip Schlosberg for always encouraging my scientific curiosity. I'm forever thankful for all you have sacrificed to encourage my education. I am blessed to have such an amazing group of close friends and new family: Matthew Anderson, Dr. Bryan and Carly Dannowitz, John and Lisa Earls, Kevin and Patricia Egan, Ashley Veljko, Tony Cucinello, John Krillich, Ricky Scampini, Monica Schlaich, Alby Sanz-Guerrero, Joey Martis, Karina, Katia, and Wayne Dabrowski. I could not have made it through these past years without your friendship and love.

Finally, I sincerely thank my wife Adriana for being most patient and supportive spouse I could ask for through these past years. She taught me to trust and believe in myself, and my life has been enriched beyond imagine because of it. Above all, I give thanks and praise to God for His divine inspiration and motivation for this work and throughout my life.

Christopher Schlosberg

Washington University in St. Louis

May 2017

ABSTRACT OF THE DISSERTATION

Modeling complex patterns of differential DNA methylation that associate with gene expression changes

By

Christopher Schlosberg

Doctor of Philosophy in Biology and Biomedical Sciences

Computational and Systems Biology

Washington University in St. Louis, 2017

Assistant Professor John R. Edwards, Chair

Gene expression is driven by specific combinations of transcription factors binding to regulatory sequences to define cell type expression profiles. Changes in DNA sequence alter transcription factor binding affinities and gene expression, and DNA methylation is an additional source of variation that is maintained throughout cellular division. Numerous genomic studies are underway to determine which genes are abnormally regulated by DNA methylation in disease. However, we have a poor understanding of how disease-specific methylation variation affects expression. Global DNA demethylation agents have been clinically approved for use in cancer, which has spurred interest in identifying genes which would be most susceptible for targeted demethylation therapies. In this work, I developed multiple tools to increase our knowledge about the relationship between methylation and gene expression in both tissue specificity and disease.

I first developed a computational strategy to identify amplifications and deletions from restriction enzyme-based methylation datasets. In a model of endocrine therapy resistant breast

cancer, I identify *ESR1* as the most amplified genomic region in response to estrogen deprivation. I develop a qPCR-based assay to probe the amplification in cell lines, formalin-fixed paraffin embedded samples, patient tumors, and xenograft samples. This data is consistent with the hypothesis that in a subset of patients, the *ESR1* amplification results in increased levels of ERα. These are produced in response to estrogen deprivation to sensitize breast cancer to low available quantities of estrogen for cellular growth.

Next, to explain specific variation in methylation that associates with expression change in both disease and tissue-specificity, I developed an integrative analysis tool, Methylation-based Gene Expression Classification (ME-Class). This model captures the complexity of methylation changes around a gene promoter. Using whole-genome bisulfite sequencing and RNA-seq datasets from different tissue samples, ME-Class significantly outperforms published methods using methylation to predict differential gene expression change. To demonstrate its utility, I used ME-Class to analyze different hematopoietic cell types, and identified that expression-associated methylation changes were predominantly found when comparing cells from distantly related lineages, implying that changes in the cell's transcriptional program precede associated methylation changes. Training ME-Class on normal-tumor pairs indicated that cancer-specific expression-associated methylation changes differ from tissue-specific changes. I further show that ME-Class can detect functionally relevant cancer-specific, expression-associated methylation changes that are reversed upon the removal of methylation in a model of colon cancer.

Lastly, I extended ME-Class to incorporate 5-hydroxymethylcytosine and uncovered gene regulatory logic involving 5hmC and 5mC in mammalian development and disease. As more large-scale, genome-wide, differential DNA methylation studies become available, tools such as

ME-class will prove invaluable to understand how specific methylation changes affect transcription. Our results show this toolset can identify genes that are dysregulated by methylation in disease, and could be used to facilitate the identification of patients who may benefit from clinically-approved demethylating therapeutics.

Chapter 1. Introduction

1.1 Gene Regulation and Epigenetics

Transcription factors (TFs) control gene expression to determine cell type specificity. For transcription to occur, TFs recruit RNA polymerase II, which must be able to access a gene's promoter sequence. DNA is spatially compacted by wrapping around histone octamers called nucleosomes. DNA methylation impacts chromatin structure in the contexts of X-inactivation, imprinting, and transposable element silencing. However, it is unclear what role DNA methylation plays outside these contexts. Multiple international consortiums including the Roadmap Epigenome Project and Blueprint Epigenome Project have mapped DNA methylation, chromatin modifications, and transcriptional activity in primarily human tissue and tumor samples (1, 2). These studies provide a fundamental resource upon which to build tools to identify which genes have their expression controlled by DNA methylation.

1.2 DNA Methylation

DNA methylation is a reversible chemical addition of a methyl group (CH_3) occurring at the 5' position in the pyrimidine cytosine. DNA methylation is also commonly referred to as 5mC and will be referred to interchangeably throughout this thesis. DNA methylation is heritable and stable through progeny, and thought to serve as a form of cellular memory to increase the information content of the genome (3). Since mammalian DNA methylation was initially discovered in 1948 (4), its function has been frequently debated (5).

In the human genome, DNA methylation primarily exists at the ~28M strand symmetric CG dinucleotide sites (CpGs), of which roughly 70% are normally methylated in somatic tissues (6). Promoter specific CpG density can also be separated into two distinct classes of having high- or

low-CpG density, depending on each promoter's observed/expected CpG density (7). High CpG density regions are commonly referred to as CpG Islands and are bounded by regions of decreased CpG density called CpG Island Shores (2-4kb from CpG Island). Adjacent to these CpG Island Shores are regions of low CpG density called CpG Island Shelves (>4kb from CpG Island) (8). Non-CpG methylation is lowly represented in most differentiated tissues, comprising <1% of all methylation. However, non-CpG methylation represents up to ~15% of mammalian embryonic stem cell (HSCs) and brain tissue methylation (6). Genome-wide, single base pair resolution maps of DNA methylation can be obtained with Whole Genome Bisulfite Sequencing (WGBS). In this sequencing technology, unmethylated cytosines are converted to uracils with sodium bisulfite while methylated cytosines are protected from conversion and followed by next generation high throughput sequencing (6).

Proper establishment of DNA methylation is essential for development, and maintenance of these patterns are stable under normal conditions. In mammalian development, primordial germ cells undergo a global wave of DNA demethylation (9), resulting in a decrease of 30% of somatic methylation levels (10), which are subsequently recovered to endogenous levels by 7W in humans and E16.5 in mice (11). *De novo* methylation at unmethylated CpGs occurs via DNA methyltransferases: DNMT3A or DNMT3B (12). Maintenance DNA methyltransferase DMNT1 binds to CpG hemimethylation and is responsible for proper of establishment of DNA methylation patterns after DNA replication (13). DMNT3L is a catalytically inactive enzyme that functions to increase the methyltransferase activity of DMNT3A/B (14). As a reader of DNA methylation, (methyl-CpG-binding protein 2) MeCP2 is an X-linked enzyme that binds to methylated CpGs thought to mediate transcriptional repression involving histone deacetylation

(15). In neurons, MeCP2 has been associated with activity and repression in a variety of genomic targets (16). MeCP2 is part of a larger family of methyl binding proteins which contain a methyl-CpG binding domain (MBD). MBD1, MBD2, and MBD4 have been identified to bind with methylated DNA and facilitate transcriptional repression (17).

Genetic experiments have elucidated the importance of DNA methylation enzymes. Inactivation of both DNA methylation writers DNMT3A/B blocks the enzymes' inherent *de novo* methylation functionality and results in embryonic lethality (12). DMNT3A KO mice with transplanted hematopoietic stem cells (HSCs) demonstrated predisposition to multiple hematological malignancies (18), and conditional KO of DMNT1 heavily redistributes HSC fate towards the myeloid lineage (19, 20). Introduction of germline mutations of DMNT1 resulted in global demethylation and embryonic lethality (21). DMNT1, DMNT3A, and DMNT3B knockout experiments in mice demonstrate that establishing DNA methylation patterns in embryogenesis is essential.

1.3 DNA Demethylation and 5-hydroxymethylcytosine

DNA demethylation may occur in a passive or active manner. Passive demethylation corresponds with the loss of DNA methylation due to incomplete copying of DNA methylation without the maintenance DNA methyltransferase DMNT1 (13). Passive demethylation requires at least 2 cell divisions. Active demethylation is an open area of research, however, there exist demethylation enzymes that participate in this process. The Ten-Eleven Translocation (TET) family (1/2/3) of enzymes are CXXC-domain containing dioxygenases responsible for oxidative conversion of 5-methylcytosine to 5-hydroxymethylcytosine (5hmC) (22, 23). TET enzymes also

can convert 5hmC to intermediates 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC), however at low detectable levels in the genome (24). 5hmC is not as ubiquitous in mammalian tissue types like 5mC, however, it exists in measureable quantities in embryonic stem cells, brain tissues (e.g. neurons), and the liver, lung, and placenta (24). These low levels of 5hmC are thought to exist because 5hmC is a demethylation intermediate since it is not maintained through mitosis. DNA methylation reader MeCP2 can also bind to 5hmC in neurons and associates with active gene expression (25). TET1- and TET2-KO mice are viable, although TET1 mice exhibit growth inhibition, decreased levels of 5mC, and increased levels of 5hmC (26). TET1-, TET2-, TET3-triple KO mice displayed impaired differentiation and embryonic development and significant promoter hypermethylation (27). Genome-wide, single base pair resolution maps of 5hmC can be obtained through either oxBS-seq (oxidative bisulfite sequencing) through the oxidation of 5hmC to 5fC via $KRuO_4$ and two bisulfite sequencing experiments or TAB-seq (TET1-assisted bisulfite sequencing) by protecting 5hmC groups from TET oxidation through glucosylation (28).

1.4 Genomic Organization of DNA Methylation Patterns

DNA methylation enzymes lay the foundation of DNA methylation patterns throughout the genome. Globally, individual CpG DNA methylation levels are highly autocorrelated up to distances of 1kb. DNA methylation shares 5-15% similarity to other somatic tissue types and up to 20% similarity to germ line samples. Greater similarity between identical tissue samples from different individuals than between somatic tissue types (29). Inter-individual human variation of DNA methylation is relatively constant at ~80% of CpGs in the genome. Dynamic variation exists at approximately ~20% of CpGs mostly surrounding low and intermediate CpG density, tissues specific promoters and enhancers (30).

DNA methylation levels are highly dependent on their genomic context. Throughout the genome, ~70-80% of all CpGs are stably methylated in somatic tissues (10). In gene-poor regions, high DNA methylation exists in regions of pericentric heterochromatin, which is essential for maintaining genomic integrity (31). Low methylated regions called DNA methylation valleys (DMVs) ranging from 5-60kb in length that are enriched for early developmental regulatory genes (32). Long range hypomethylated domains that are associated with nuclear lamina attachment and late replication genes (33). Partially Methylated Domains (PMDs) are large contiguous regions of decreased methylation (~70% over ~150kb) often containing downregulated genes (6). CpG methylation may also belong to a class of Intermediate Methylation (IM), enriched for enhancers, exons and DNaseI hypersensitive sites (34), however, it is unclear whether IM sites and PMDs derive from clonal heterogeneity. Enhancers often display tissue specific methylation in the mouse (35) and human genome (1).

In genome-wide averages of gene annotations in normal tissues, the transcription start site (TSS) is typically unmethylated and there is an increase in gene body methylation throughout the length of the gene until the transcription end site (TES), where methylation levels return to intergenic levels (1, 36). DNA methylation also exists throughout the gene body, where it is thought to decrease the elongation capability of PolII transcriptional machinery (37).

The genomic distribution of 5hmC patterns in the genome is limited due to its relative scarcity in differentiated tissues. In mammalian ESCs, 5hmC is enriched within gene bodies and enhancers (38). 5hmC has been mapped in murine and human brain development and TET2 is enriched at

demethylated CpGs (39). Promoter specific analyses have shown conflicting results as to the effect of TET1/2 KOs on 5hmC levels in mouse models (40, 41).

1.5 Histone Modifications

DNA methylation works in concert with histone modifications to provide accessibility for RNA Polymerase II (PolII) necessary for active gene expression (42). Histone modifications also interact with the transcriptional machinery to affect expression levels. While there is a lack of evidence to suggest a strictly deterministic histone code hypothesis (43, 44), many patterns of histone modifications affecting gene expression through chromatin accessibility are well characterized. Punctate patterns of H3K4me3 and H3K9ac is characterized to associate with active genes. Broad peaks of H3K36me3 in the gene body and near the TES has also been associated with active genes. Punctate peaks of H3K27me3 near the TSS is a primary characteristic of inactive genes (45). However, when H3K27me3 and H3K4me3 are found in combination near the TSS, the gene is known to be in poised or bivalent chromatin configuration, lending to either active or inactive gene expression (46).

DNA methylation and H3K9me3 interact by recruiting DMNT3A/B for *de novo* methylation via a relationship mediated by SET domain-containing histone methyltransferase enzymes (SUV39H1 and SUV39H2) in silencing satellite DNA sequences (42) or via G9A to silence pluripotency genes (47). For active genes, H3K4me3 (at the TSS) and H3K36me3 (in the gene body) are mutually exclusive with DMNT3A/B/L binding and unmethylated CpGs near the promoter (48).

1.6 DNA methylation in normal function

DNA methylation has well defined roles in silencing of gene expression in imprinted genes (49), inactivation of the X chromosome in female mammals (50), and silencing of transposons (51). DNA methylation maintains X-inactivation during normal development, where a single copy of the female mammalian X chromosome is permanently condensed (52). DNA methylation also plays a role in imprinting, such as the parent-of-origin allele specific gene expression of the *H19/IGF2* locus (49). During hematopoiesis, DNA methylation has been shown to be involved in myeloid-lymphoid lineage commitment (53) and characteristic TF binding of *GATA1* and *TAL1* in methylated lymphoid progenitors (54). DNA methylation also serves to silence transposable elements (51, 55) and to promote genome stability (56). These observations were shown to induce transcription of transpositionally active intracisternal A particles (IAP) upon DMNT1 deletions in mouse embryos (51). Alu and L1 retroelements contained within human gene promoters were also shown to elevate repeat transcripts upon treatment with demethylating agents (57, 58). Recent studies have provided experimental evidence that induced DNA methylation (via dCas9-DMNT) can silence expression (59, 60), and also that targeted demethylation (via dCas9-TET) of promoters and enhancers can reprogram expression profiles (61).

1.7 DNA methylation and disease

DNA methylation is involved in multiple developmental human diseases. Prader-Willi syndrome (PWS) results in severe developmental disabilities due to the loss of paternal expression in the PWS region on chromosome 15, which can be detected by abnormal, imprinted methylation patterns (62). Beckwith-Wiedemann syndrome occurs when the loss of an imprinting mechanism

results in the overexpression of IGF2 on chromosome 11, which leads to dysregulated growth (63). Compromised immune ability and chromosomal instability are characteristic features of Immunodeficiency–centromeric instability–facial anomalies (ICF) syndrome caused by autosomal recessive inheritance of a mutation in the DNA methylation writer DMNT3B (64). Rett syndrome is a developmental brain disorder caused by mutations in the X-linked DNA methylation reader MECP2 (65).

In human cancers, DNA methylation has been shown a valuable source of added information as a biomarker where single changes of DNA methylation are as frequent, if not are more frequent than the number of somatic mutations (66). DNA methylation is primarily involved in silencing tumor suppressor genes through targeted promoter hypermethylation (67) in a genomic background of global hypomethylation (33, 68). This targeted *de novo* hypermethylation primarily occurs at promoters associated with CpG Islands. CpG Island Shores have been implicated in their correlation with gene expression in colon cancer (69). CpG Island Shelves show a similar level of correlation with gene expression as CpG Island Shores in chronic lymphocytic leukemia (CLL) (70). Demethylating agents have bene created to reactivate tumor suppressor genes by removing promoter methylation in tumors (71). Indeed, demethylating agents 5-aza-2'-deoxycytidine (Decitabine) and 5-azacitidine (Azacytidine) have been clinically approved for the treatment of Myelodysplastic Syndrome (MDS), a precursor to leukemia, and Acute Myeloid Leukemia (AML), respectively. Enhancers (as defined by p300 and H3K27ac) are also preferentially hypermethylated and correlated gene expression with in colon cancer (33). Loss of function and overexpression mutations exist in DNMT3A/B and DNMT1, resulting in poor clinical outcomes for patients with AML, MDS, and colorectal cancer (72). Mutations in

TET2 in human breast, liver, lung, pancreatic and prostate cancer samples result in decreased *TET2* expression and decreased levels of 5hmC (73).

1.8 Copy Number Variants

Many complex phenotypes and human diseases are caused by deletions, amplifications, and translocations of genomic DNA. These Copy Number Variants (CNVs) have lasting implications on gene transcription. In a study of HapMap populations, CNVs (17.7%) occur less frequently than Single Nucleotide Polymorphisms (SNPs) (83.6%), however, they both have a complementary impact on gene expression (74). In cancer, translocations can result in chimeric gene products that can have resultant increased expression to aberrantly increase the expression of malformed protein products (75). Deletions can result in a dosage loss of important tumor suppressors, while amplifications can result in the overproduction of selective oncogenes (76). To identify these modifications, array comparative genomic hybridization (aCGH) is a longstanding assay technique to detect amplifications and deletion at megabase resolution (77), and whole genome sequencing has increased the ability to more precisely detect these amplifications with base pair resolution (78, 79). DNA methylation array (80) and affinity-based sequencing (81) assays have been used for the detection of CNVs. DNA methylation data is inherently impacted by CNVs and array probes must be corrected for copy number before correlation with gene expression (82). However, DNA methylation has been successfully used to detect CNVs (83). The coordinated assay of CNVs and DNA methylation is also preferred because it would allow us to identify which modification contributes to gene expression change (84). In addition, SNPs have been involved in association with gene expression showing high expression genes having relatively low methylation, and low expression genes having an

increased level of methylation variability (82, 85). The discovery of these modifications has led to concerted effort to categorize human genetic and epigenetic variation in normal tissues and disease.

1.9 Associations between gene expression change and methylation

Multiple international collaborations are assaying variation in DNA methylation and chromatin modifications over multiple human tissues (Roadmap Epigenomics Project) (1) and differences in hematopoietic differentiation (Blueprint Epigenome) (2). These studies characterize methylation levels throughout the genome by examining the variation in single samples. Further analysis has shown that DNA methylation has tissue specific patterns that are not restricted to methylation in CpG islands alone (86).

In a study of 33 normal tissue and cell line WGBS samples, low expressing genes show average DNA methylation is predominately characterized by increased methylation at the TSS (~30-40%) and decreased methylation throughout the gene body (~70-75%). Conversely, high expressing genes show decreased methylation at the TSS (~10-20%) and increased methylation throughout the gene body (~75-80%) (1). DNA methylation has also been reduced to multiple averaged values across annotated gene elements (upstream, exon, intron, and downstream bins) to associate with expression levels showing similar results in single samples (87). While these studies summarize DNA methylation levels in individual samples, researchers often examine differential methylation, such as the difference between normal and diseased states, and associate with expression changes to identify functional differences.

Typically, promoters are labeled as either methylated and silenced or unmethylated and potentially active (10, 88). Although most analysis techniques rely upon this simple binary characterization (89), studies that model methylation using a single window (SW) of ~2kb around the promoter region find only modest negative correlations with expression levels (6, 82, 90, 91).

The most common current approach to associate DNA methylation and expression change is to first identify differentially methylated regions (DMRs) and then associate them with nearby genes. Numerous statistical tools have been developed to identify DMRs (reviewed in (89)). Generally, DMRs are found by segmenting the genome into single CpGs or larger equally spaced regions and statistical significance is assigned to each region (89). Biological insight is gained when known genomic regulatory elements are associated with DMRs within a certain distance. However, DMR methods rely on a set of arbitrarily defined thresholds for the size and number of CpGs to include in the DMR. It is often recommended to adjust these parameters for each individual dataset, since the choice of these parameters has substantial implications in the numbers of DMRs identified and putatively regulated genes. A standard integrated analysis of DNA methylation and gene expression starts by first defining the most variable DMR regions. These regions are then overlapped with the largest expression changes, and then top candidates with an inverse correlation are chosen (92). Indeed, studies often find only weak correlations between DMRs near gene promoters and differential gene expression (93-95).

One possible reason both the SW and DMR methods fail to find strong association between differential methylation and expression is they reduce DNA methylation to a single differential

value without considering the local context of these changes. As a result, more differential methylation associations with expression found by DMRs are identified within enhancers rather than near promoters (30). Using a discovery approach (96), we have previously shown the importance of capturing all methylation changes around the TSS to find patterns of methylation change that associate with expression changes in a model of DMNT inhibition in AML (97) and cellular senescence (98). As detailed in the explanation of ME-Class in Chapter 3, we incorporate all DNA methylation changes within a 10kb centered window of the TSS to make minimal assumptions about how methylation changes are predictive of transcriptional changes. The purpose of ME-Class is to provide stronger evidence for methylation associated expression changes by training from examples of methylation change to set model parameters then predict in unseen datasets.

1.10 Research Motivation

Various international consortiums and individual labs have collected epigenome-wide maps of methylation and transcriptional data on multiple cell types and disease states. It is imperative that we convert this wealth of resources into fundamental knowledge about DNA methylation associated gene expression changes. We hypothesize that complex patterns of DNA methylation are necessary to understand the relationship between DNA methylation and gene expression. In Chapter 2, we first identified and validated targets of copy number variants using a model of endocrine therapy resistant breast cancer and patient xenografts. In Chapter 3, we develop a computational tool to classify differential gene expression from genome wide DNA methylation patterns. Our tool, ME-Class (Methylation-based Expression Classification), enumerates putatively functional patterns of DNA methylation in tissue and cancer samples. Using this tool,

we identified that 3’ TSS proximal changes are most important, but not sufficient, for expression classification in comparisons of 17 different tissue samples from the Roadmap Epigenomics Project. We also validated our method using 32 datasets from different hematopoietic cell types from the Blueprint Epigenome project supporting the hypothesis that methylation changes precede transcriptional changes in normal blood development. ME-Class was also used to identify a subset of genes that were preferentially upregulated upon demethylation in a model of colon cancer. In Chapter 4, we extended the capability of ME-Class to include 5hmC in the prediction of differential expression classification. We identify frequencies of patterns of 5mC and 5hmC corresponding with differential expression classes in a model of mouse brain development and of human tissue- and cancer-specificity. In Chapter 5, we outline our hope for ME-Class and its future extensions. In the future, we argue that ME-Class can be useful in identifying suitable patients for clinically approved demethylating therapeutics.

Chapter 2. Endocrine-Therapy-Resistant *ESR1* Gene Amplification Revealed by Genomic Characterization of Breast-Cancer-Derived Xenografts

---

This chapter is adapted from Li S, Shen D, Shao J, Crowder R, Liu W, Prat A, He X, Liu S, Hoog J, Lu C, Ding L, Griffith OL, Miller C, Larson D, Fulton RS, Harrison M, Mooney T, McMichael JF, Luo J, Tao Y, Goncalves R, **Schlosberg CE**, Hiken JF, Saied L, Sanchez C, Giuntoli T, Bumb C, Cooper C, Kitchens RT, Lin A, Phommaly C, Davies SR, Zhang J, Kavuri MS, McEachern D, Dong YY, Ma C, Pluard T, Naughton M, Bose R, Suresh R, McDowell R, Michel L, Aft R, Gillanders W, DeSchryver K, Wilson RK, Wang S, Mills GB, Gonzalez-Angulo A, Edwards JR, Maher C, Perou CM, Mardis ER, Ellis MJ. Endocrine-therapy-resistant *ESR1* variants revealed by genomic characterization of breast-cancer-derived xenografts. *Cell Rep.* 2013 Sep 26;4(6):1116-30.

2.1 Introduction

Breast cancer is the most commonly diagnosed cancer and results in the second highest level of cancer mortality in the United States (99). Estrogen Receptor (ER), Progesterone Receptor (PR), and HER2 expression are among common molecular markers to subtype breast cancer. Subtyping of patients is commonly based these molecular markers to define luminal A (ER+/PR+/HER2-), luminal B (ER+/PR+), HER2-enriched (ER-/PR-/HER2+), basal-like (ER-/PR-/HER2-), and normal-like (ER+/PR+/HER2-) subtypes (100). Based on this work, the PAM50 expression gene signature was developed to identify subtypes of breast cancer (101) and help stratify patients according to their clinical outcomes (102, 103) and responses to chemotherapy (104). Even though ER+ breast cancer cases comprise ~70% breast cancer patients, ER+ breast cancers show a high degree of heterogeneity (105), and almost a third of ER+ breast cancer patients experience recurrence after treatment (106). While gene expression profile methods have been successful in categorizing patients for treatment regimes, these gene expression markers can be improved to better predict response to treatment. The ERα protein product is encoded by the *ESR1* gene. ERα functions as a transcription factor and forms a homodimer or heterodimer with the protein product of *ESR2* (ERβ) (107). Clinical therapeutics, either through ER antagonism (tamoxifen) or aromatase inhibition (letrozole), have been developed to inhibit estrogen production and are commonly referred to as estrogen deprivation therapy. However, many patients experience resistance after extended exposure to these therapies (108). Cell line models have been created to understand long term estrogen deprivation (LTED) therapies (109). LTED cell lines re-exposed to estradiol counterintuitively slows cellular growth to a state defined as LTED-Recovered (LTED-R). To develop new therapies, we need a better understanding of the genetic and epigenetic mechanisms of resistance.

Copy number variants (CNVs), such as amplifications and deletions, are inherently common to ER+ breast cancer patients and have been shown to be indicative of response to aromatase inhibition (110). Structural variants, such as translocations, have also been identified in breast cancer cells (111). Deletions, insertions, translocations can disrupt transcription, while amplifications often increase the transcription. Translocations can alter regulatory environments and place two distal, functional elements in close proximity if the translocation does not interrupt the reading frame (112). While CNVs (~13% of all somatic cancer variation) are less frequent than Single Nucleotide Polymorphism (SNPs), SNPs show low commonality between somatic cancers and have many private point mutations (113). CNVs often are more indicative of the underlying etiology of the disease rather than SNPs (114), and show a higher locus specific genome arrangement rate than point mutations (115). In a recent study of somatic CNVs in multiple cancers in TCGA, breast cancer patients with normal ploidy displayed focal amplifications at a frequency of ~20% and deletions at a frequency of ~15% (116). In a separate study, ERBB2, MYC, FGFR1/ZNF703, CCND1, and ZNF217 were identified as most common genes with copy number changes in ER+ breast cancer patients (110). One controversial, low frequency amplification surrounds the *ESR1* transcript 1 alternative promoter region, which has been observed to predict response to estradiol treatment (117). There exists an open debate as to the actual frequency of this amplification due to inconsistent technical considerations of the methods used for detection, however, estimates vary from ~1-33% of ER+ breast cancer patients harboring this CNV (118-121).

DNA methylation is also significantly altered in breast cancer and is defined by global hypomethylation and targeted promoter hypermethylation (122). Endocrine therapy resistance is often characterized by hypomethylation and oncogene activation. Genome-wide sequencing has previously been inhibited by cost. However, Methyl-MAPS, a single base pair resolution restriction enzyme based DNA methylation sequencing assay (36), is an advantageous, opportunistic solution for the analysis of CNVs during breast cancer models of LTED. Here, we conduct a screen for CNVs in multiple models of endocrine therapy resistant breast cancer. To effectively model an advanced stage human breast tumor in mice, patient derived xenografts (PDXs) or "Washington University Human in Mouse" (WHIM) lines were created. Using PDXs and cell line models, we examine our ability to use Methyl-MAPS data for the prediction of CNVs and to evaluate the presence of the *ESR1* amplification.

2.2 Methods

2.2.1 Breast Cancer Cell Lines and Patient Samples

ER+ Parental (no estrogen deprivation), long term estrogen deprived (LTED) cell lines (MCF7, T47D, MDA415, HCC1428), and LTED recovered (MCF7 LTED-R) are prepared as described in Sanchez *et al.* (109). The MCF7 LTED-R cell line refers to a MCF7 LTED cells exposed to estradiol, which counterintuitively slows cellular growth. Throughout the study, MDA415 corresponds to cell line MDA-MB-415. Experimental design for patient tumor samples and xenografts (WHIMs) are as described in Li *et al.* (123).

2.2.2 Identification of CNVs and SVs from Methyl-MAPS

We initially mapped cell line data with BWA v0.5.10 with <=2 gaps per read, <=4 total mismatches, <=1 gap opening, and a colorspace index to map Methyl-MAPS read pairs. Deletions and amplifications are detectable by read depth information, while translocations and inversions are often best detected by read pair information (112). To identify amplifications and deletions, we developed a sliding window method. This sliding window method consisted of a 200bp windows overlapping by 50bp across the summed and max-normalized methylated (McrBC) and unmethylated (RE) genomic fractions. We then applied a genomic segmentation algorithm, Genome Alternation Detection Analysis (GADA) (124), to predict CNVs (with parameters "-a 0.8 -T 5 -M 3 -s -0.2 -b 1 -c"). CNVs must overlap by ≥50bp with Aguilar et al. (117) to be counted as shared genomic gains or losses (Supplementary Figure 2.2). We used SVDetect (125) to identify structural variants based on strand, order, insert size (400bp), and tile filtering and >=2 discordantly mapped reads in the Methyl-MAPS libraries. Structural variants were visualized using Circos. We also used Trinity (126) to confirm positive gene fusions from translocations predicted by SVDetect from matched RNA-seq data.

2.2.3 RNA-seq analysis

HTSeq-count was used for count quantification of mapped RNA-seq reads (127). Differentially expressed genes were identified with edgeR using FDR-corrected p-value cutoff with α=0.05 (128).

2.2.4 Quantitative Real-Time PCR to Determine *ESR1* Amplification

Primers were designed using Primer3 (129) (Supplementary Table 2.1). Assays were optimized and run on a Viia 7 Real-Time PCR System (Life Technologies) using 10ng of genomic DNA for each sample according to the manufacturer's instructions. One set of primers was used for each of the control genes FAM38B and ASXL2, as described previously (121). FAM38B and ASXL2 were used as controls because they are not affected by copy number polymorphisms in http://projects.tcag.ca/variation/ nor affected by gains or losses in a breast cancer aCGH database. Three primer sets were used to interrogate the amplification at locations 1, 2, and 3 (as specified in Figure 2.1). p-values were computed by a one-way ANOVA using Dunnett's post hoc test using R. Three replicates were performed for each qPCR.

2.3 Results

2.3.1 CNVs identified in LTED models from MethylMAPS data

We first examined differentially expressed genes between MCF7, T47D, and MDA415 breast cancer cell lines and observed little common overlap between either upregulated (9 genes in triple intersection) or downregulated genes (7 genes in triple intersection) (Supplementary Figure 2.1). We sought next to identify concordant expression changes with amplifications or deletions in the breast cancer cell lines. To identify CNVs, we developed a sliding window method, max-normalized to both the methylated (McrBC) and unmethylated (RE) genomic fractions, followed by a genomic segmentation (124). We compared our method with a previous study examining MCF7-LTED with a microarray based CNV approach (117), however we identified few shared genomic amplifications (n=13) and deletions (n=17) upon LTED treatment of MCF7 (Supplementary Figure 2.2). We observed there were few concordant amplifications with

upregulation and deletions with downregulation shared amongst the breast cancer cell lines (Supplementary Figure 2.3). This is not unexpected given the lack of overlap in differentially expressed genes between the three cell lines (Supplementary Figure 2.1). However, we recognized a severe bias in normalized coverage between methylation fractions which raised concern over the false positive rate of CNV detection and so we focused our analysis on significant amplifications.

2.3.2 SVs identified in LTED models from MethylMAPS data

We next identified SVs by using ambiguously mapped, mate-paired Methyl-MAPS reads within MCF-7 Parental (Supplementary Figure 2.4). While this analysis returned many putative candidate genes for further analysis, there was a lack of concordance in comparison to a published analysis of CNVs and SVs in MCF7 from Hillmer et al. using a long pair-end sequencing approached (130). We also sought to confirm gene fusion products with RNA-seq reads produced from predicted translocations. We confirmed one fusion gene in BCAS3/BCAS4, which was not present in the Hillmer et al. analysis. While there was a large discrepancy between Hillmer et al. and our analysis, it is unclear whether this discrepancy can be explained by technical variation between using WGS versus Methyl-MAPS or by biological variation from genetic drift between each of the MCF-7 cell lines.

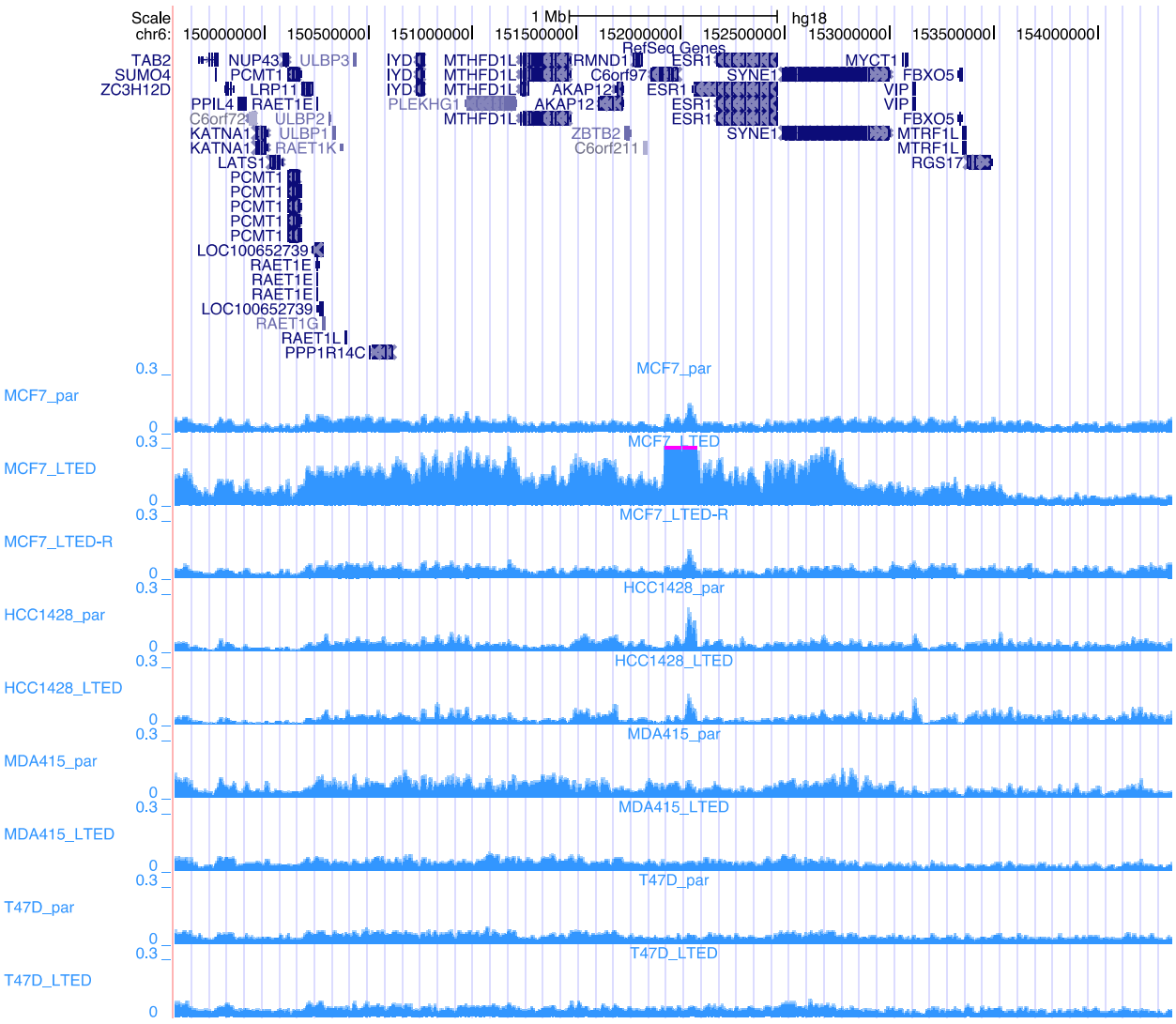


Figure 2.1 *ESR1* is amplified in ER+ cell line models of breast cancer. Shown is 200bp resolution sliding window (50bp overlapping) of max-normalized, summed Methyl-MAPS fractions.

We identified *ESR1* as the largest putative CNV with a ~20x amplification between MCF7-Parental and -LTED (Figure 2.1). There exists a focal amplification at the upstream *ESR1* alternative promoter region of transcript 1 set within a larger amplified region (~6kb) above the background coverage level. HCC1428 also displayed the same focal amplification at the

alternative promoter region, but did not show the broader amplification. T47D-Parental, T47D-LTED, MDA415-Parental, and MDA415-LTED cell lines did not display any *ESR1* amplification, even though all Parental cell lines in the study are ER+. The MCF7 LTED-R cell line showed a reduction of the *ESR1* amplification to a level similar to MCF7-Parental. These gene amplifications were consistent with increased expression in the *ESR1* transcript product in MCF7 and HCC1428.

2.3.3 *ESR1* is most amplified region in LTED model and confirmed in PDX genome

Upon identifying *ESR1* as the most amplified region in the MCF7-LTED and HCC1428-LTED genomes, we sought to examine whether this amplification existed in a xenograft model of breast cancer. We created a qPCR assay for genomic DNA using separate probes in the affected promoter and enhancer region of *ESR1* in our cell line models, an advanced stage patient sample, and the patient's mouse-human xenograft (WHIM16) sample. We visualized the gene amplification with whole genome sequencing (WGS) data across the *ESR1* promoter and coding region in WHIM16 (Figure 2.2a).

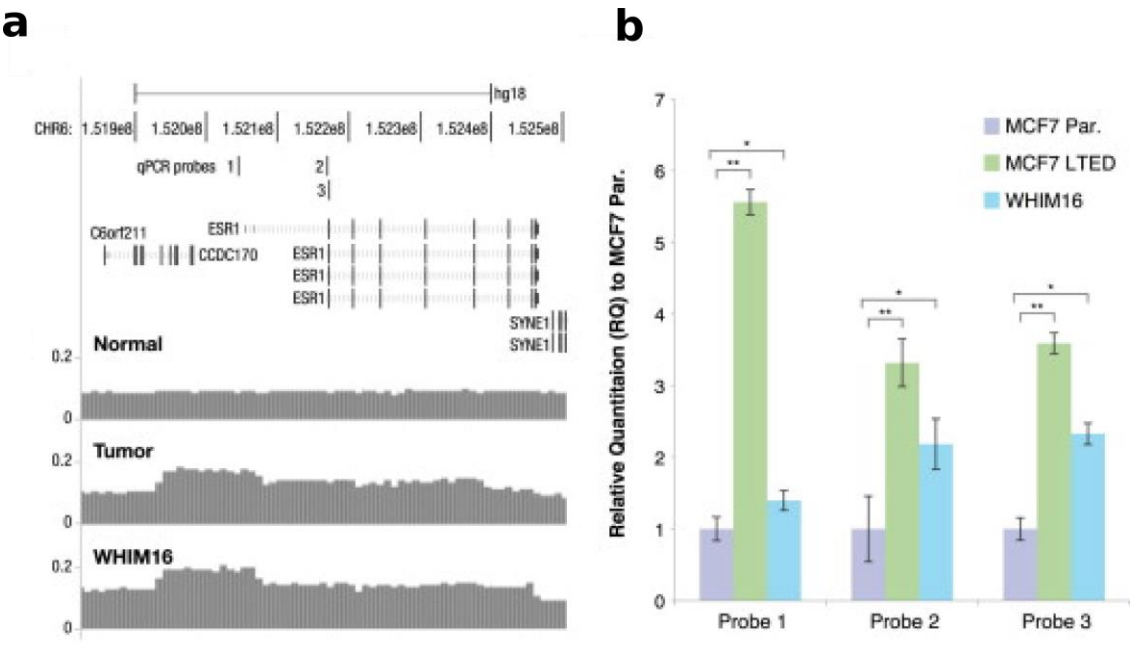


Figure 2.2 *ESR1* is a confirmed amplification in ER+ patient tumor and WHIM16 cells. a) WHIM16 and the originating tumor harbor amplification of the *ESR1* gene that extends from the promoter region throughout the coding sequence that was mapped using read counts obtained with WGS. b) qPCR on genomic DNA using three separate probes was used to confirm gene amplification in WHIM16 cells. The negative control was MCF7-Parental cells and the positive control is the MCF7-LTED displaying the *ESR1* gene amplification. qPCR results were normalized relative to MCF7-Parental. The positions of probes 1, 2, and 3 are displayed in (a). Error bars are ±1 SD of the mean relative quantification (RQ); *$p < 0.05$, ** is $p < 0.01$.

The *ESR1* amplification exists in both patient tumor and WHIM16 corresponds with high levels of ERα protein (123). In this experiment, MCF7-parental genomic DNA was used as a nonamplified *ESR1* control based on Methyl-MAPS (Figure 2.2b). These data suggest that *ESR1* amplification is an adaptation to estrogen deprivation. We examined the effect of being able to detect CNVs in formalin-fixed paraffin embedded (FFPE) samples, as patient tumors are often

preserved with this method (131). To recapitulate genomic DNA in an FFPE sample, we sheared genomic DNA using probe sonication to shear MCF7-LTED data to 10ng. In this experiment, T47D genomic DNA was used as a nonamplified *ESR1* control, and we still observed the *ESR1* amplification in the sonicated MCF7-LTED samples as well as in WHIM16. (Supplementary Figure 2.5)

2.4 Discussion

We sought to conduct a screen for CNVs from methylation data in endocrine resistant breast cancer models, and identified *ESR1* as the most amplified genomic region. To obtain a more accurate measure of methylation levels, a methylation estimation model could be built for Methyl-MAPS data. This model should be tested in normal ploidy samples with WGBS validation given the prevalence of whole genome duplication in breast cancer (116). While our estimates for CNVs and SVs from methylation data suffer from a high false positive rate, we identified a clinically relevant amplification.

We then developed a qPCR assay with proper single copy breast cancer control primers to identify this amplification in cell lines and xenograft samples. In WHIM16, this amplified *ESR1* resulted in an increased gene product of ERα. This increased level of ERα receptor would heavily sensitize the breast cancer cells to even low amounts of estrogen after estrogen deprivation therapy. The *ESR1* amplification was also observed during a 150 day period of estrogen deprivation of MCF7 cells in Aguilar et al (117). Interestingly, the MCF7-LTED-R cell line shows a reduction in the amplification (Figure 2.1) along with a decrease in cellular growth (109). Reintroducing endocrine therapy in advanced ER+ breast cancer patients has been

effectively used after endocrine deprivation therapy (132). WHIM16 also demonstrates tumor regression after the reintroduction of endocrine therapy, implying that the *ESR1* amplification could be used as a predictive biomarker for tumor sensitivity to estradiol treatment in advanced ER+ breast cancer patients.

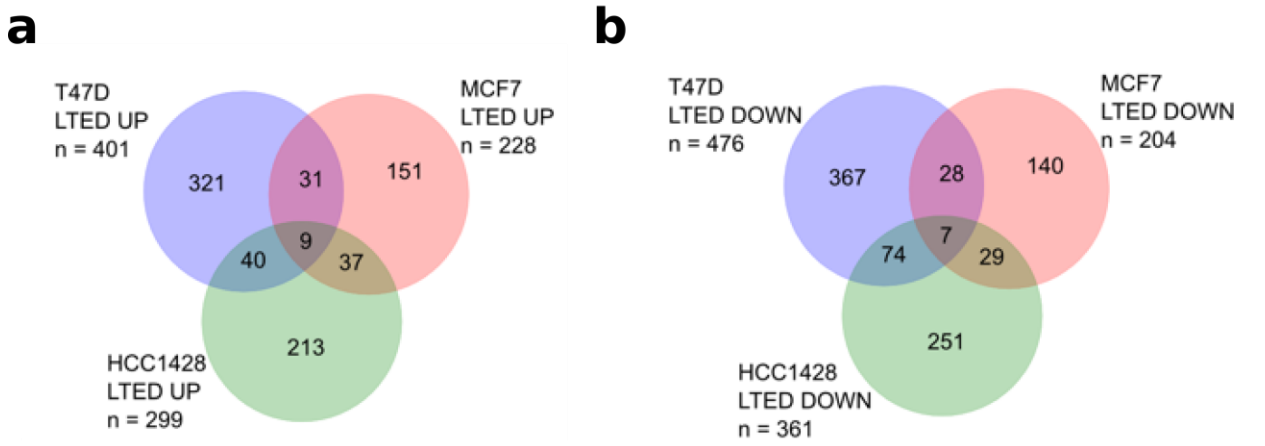
The identification of the *ESR1* amplification is still an area of active research as to the relative frequency of occurrence. The *ESR1* amplification is identified in MCF7, HCC1428, and WHIM16 and these models respond to endocrine deprivation therapy (fulvestrant). This implies that the *ESR1* amplification is important for a subset of patients and related to estrogen signaling. Initially, this amplification was identified in 20.6% of 2,000 breast cancer patients through a screen using an Affymetrix 10K SNP array (119). The large discrepancy between these studies is based on the use of the reference gene *ESR2*, which is often deleted in breast cancers. However, using *FAM38B* and *ASXL2* as controls for a qPCR-based assay, only 2.8% breast cancer tumors showed copy number gain (121). In a separate study, only 1% of breast cancer tumors identified by FISH or aCGH demonstrated the amplification. The low frequency of the *ESR1* amplification is also evidenced by a large study of ER+ breast cancer patients treated with neoadjuvant aromatase inhibitor therapy where the amplification was not identified as a common CNV (110). Even though it exists at a low frequency, we detected low levels of the *ESR1* amplification despite high levels of shearing as observed in FFPE samples. The *ESR1* amplification is not as frequent as initially reported, however, it was hoped that the amplification could identify a subset of patients that would best respond to estrogen deprivation. A recent review has shown contradictory results as to whether the *ESR1* amplification is predictive of antiestrogen response to treatment (133). However, the *ESR1* gene amplification in WHIM16 is correlated with

estradiol-induced regression and is preserved from the patient tumor sample, implying that PDXs faithfully recapitulate CNVs from the originating tumors. The presence of this amplification is consistent with *ESR1* overexpression providing an adaptive advantage for tumor growth in the absence of estrogen. The additional ERα may better scavenge for estrogen to directly function as a gene regulator of ER gene targets.

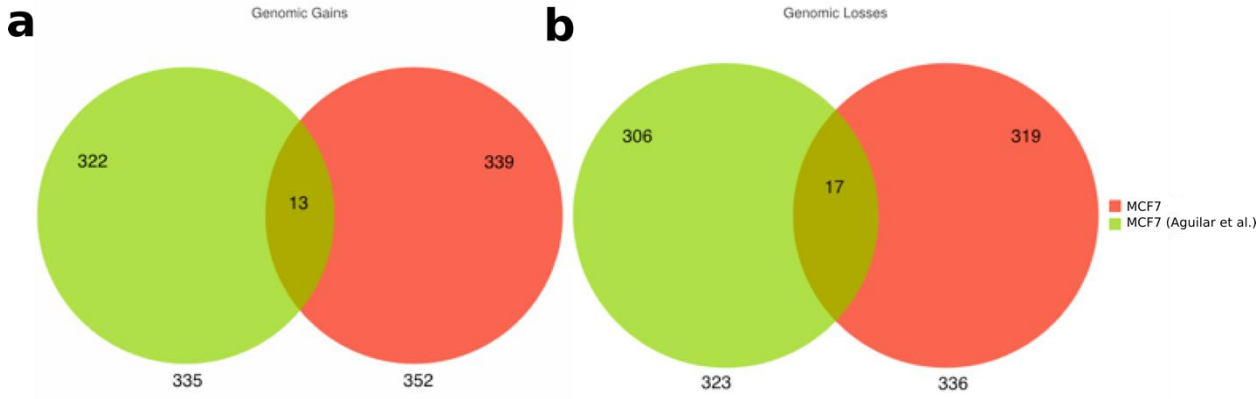
Supplementary Table 2.1 Primer sequences for qPCR control and target genes.

	Control Primers	
Gene Target (Region)	**Forward**	**Reverse**
FAM38B	5'-GAA ACC CCC TTC CTA AGC AC-3'	5'-AGC CTG CGT TCT CCA TAA GA-3'
ASXL2	5'-CAG CTT CTC ACT TGG CCT TC-3'	5'-GCT CTG CAC AGG ACA GAT CA-3'

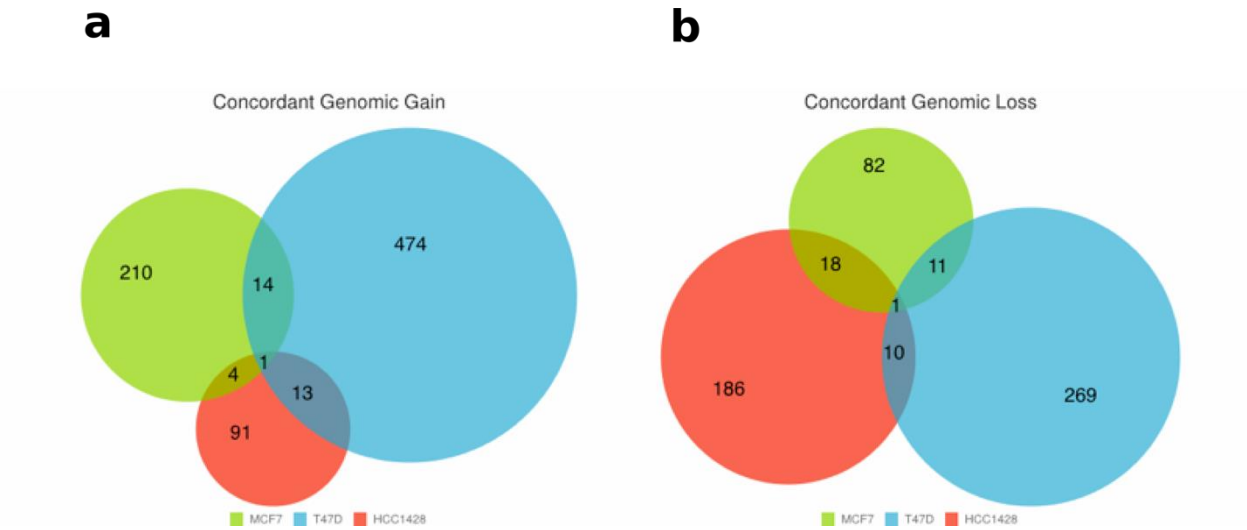
	Target Primers	
Gene Target (Region)	**Forward**	**Reverse**
ESR1 Amplification (Probe 1)	5'-AAA ATG CCT CAG GAC GAT TG-3'	5'-GCG CCT GAG AAG CTA GAG AA-3'
ESR1 Promoter (Probe 2)	5'-AAG CCC ATG GGA CAT TTC TG-3'	5'-ACA TAC CCC CAT GGA GAA CA-3'
ESR1 Exon1 (Probe 3)	5'-CCA TGA CCC TCC ACA CC-3'	5'-CTC GTT CCC TTG GAT CTG A-3'



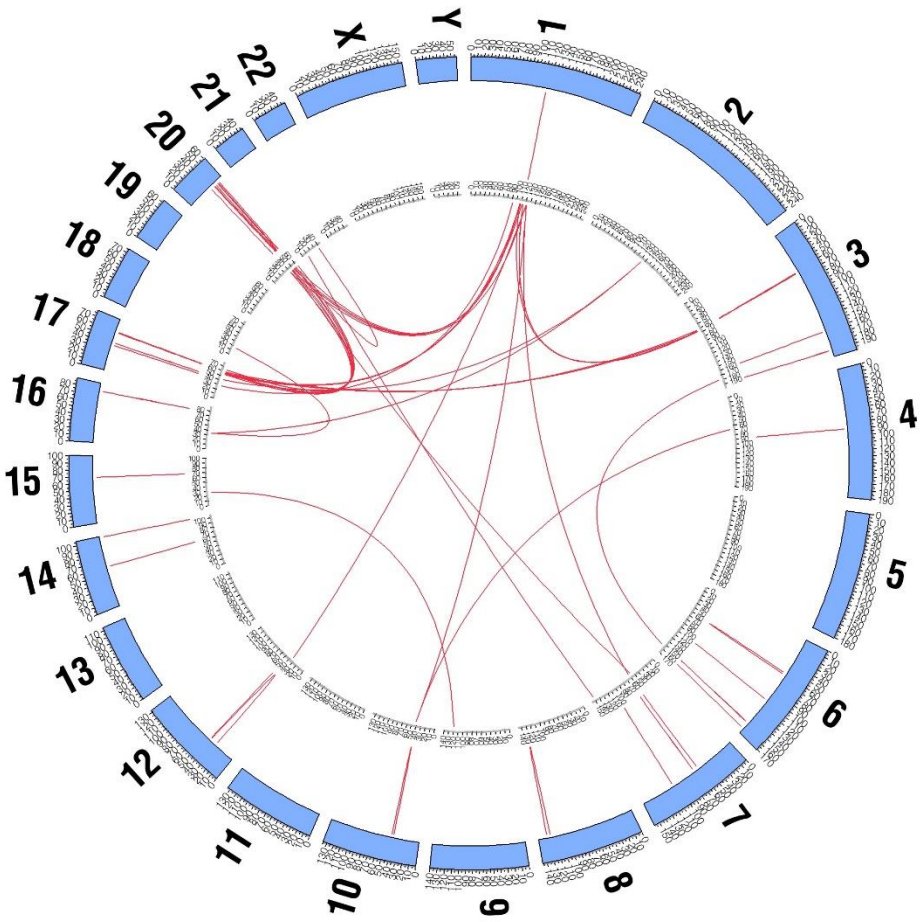
Supplemental Figure 2.1 Few shared differentially expressed genes during long term estrogen deprivation. Venn diagrams for a) upregulated and b) downregulated differentially expressed genes.



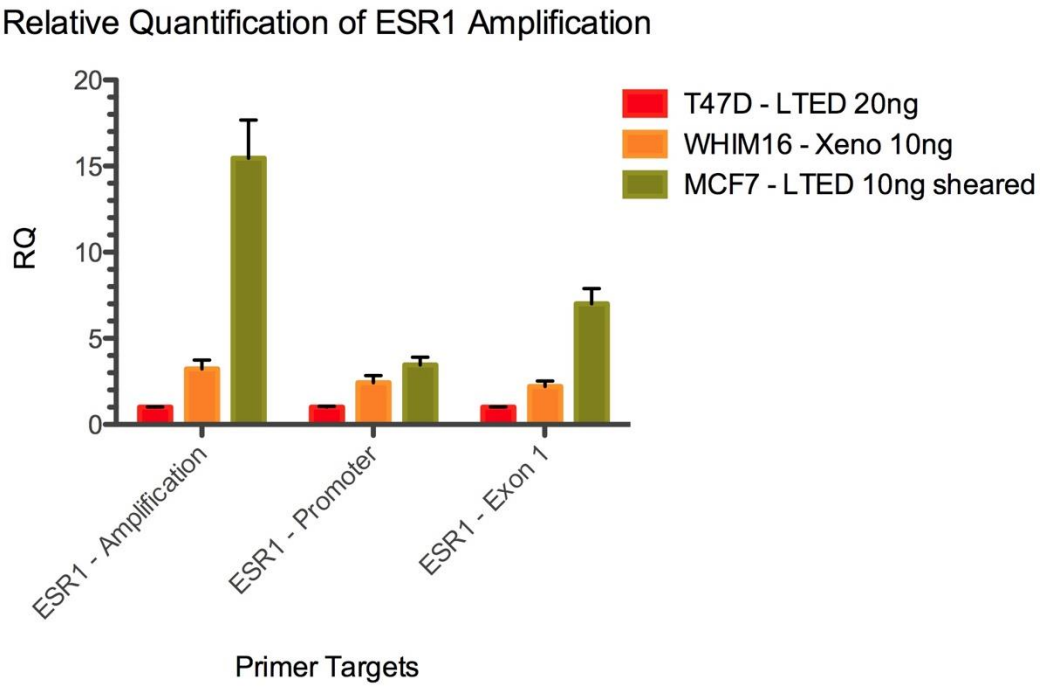
Supplementary Figure 2.2 Few Shared Genomic Amplifications and Deletions upon LTED with Aguilar et al (117). MCF7 refers to amplifications and deletions identified in this study.



Supplementary Figure 2.3 Few shared concordant differential CNV and expression changes during LTED. Identified a) amplifications and b) deletions by overlapping sliding window method.



Supplementary Figure 2.4 Circos plot showing CNVs and SVs in MCF7-Parental. Outer ring: amplifications and deletions. Inner ring: inter- and intra-chromosomal rearrangements.



Supplementary Figure 2.5 *ESR1* amplification can be identified in simulated FFPE samples. qPCR results normalized to T47D-LTED 20ng samples and reference control genes *FAM38B* and *ASXL2*.

Chapter 3. Modeling complex patterns of differential DNA methylation that associate with gene expression changes

---

This chapter is adapted from **Schlosberg CE**, VanderKraats ND, Edwards JR Modeling complex patterns of differential DNA methylation that associate with gene expression changes. *Nucleic Acids Res.* 2017 In press.

3.1 ABSTRACT

Numerous genomic studies are underway to determine which genes are abnormally regulated by DNA methylation in disease. However, we have a poor understanding of how disease-specific methylation changes affect expression. We thus developed an integrative analysis tool, Methylation-based Gene Expression Classification (ME-Class), to explain specific variation in methylation that associates with expression change. This model captures the complexity of methylation changes around a gene promoter. Using 17 whole-genome bisulfite sequencing and RNA-seq datasets from different tissues from the Roadmap Epigenomics Project, ME-Class significantly outperforms standard methods using methylation to predict differential gene expression change. To demonstrate its utility, we used ME-Class to analyze 32 datasets from different hematopoietic cell types from the Blueprint Epigenome project. Expression-associated methylation changes were predominantly found when comparing cells from distantly related lineages, implying that changes in the cell's transcriptional program precede associated methylation changes. Training ME-Class on normal-tumor pairs from TCGA indicated that cancer-specific expression-associated methylation changes differ from tissue-specific changes. We further show that ME-Class can detect functionally relevant cancer-specific, expression-associated methylation changes that are reversed upon the removal of methylation. ME-Class is thus a powerful tool to identify genes that are dysregulated by DNA methylation in disease.

3.2 INTRODUCTION

Establishment of specific patterns of DNA methylation at CG dinucleotides (CpGs) is necessary for normal development (134, 135), and aberrant methylation is frequently observed in cancer (136, 137). CpG rich-regions, often called CpG islands (CGIs), are typically unmethylated and associated with ~70% of mammalian gene promoters (6). Hypermethylation of CpG islands overlapping the transcription start site (TSS) is hypothesized to downregulate tumor suppressor genes, thus promoting tumorigenesis (138, 139). Typically, promoters are labeled as either methylated and silenced or unmethylated and potentially active based on the methylation levels near the transcription start site (TSS) (10, 88). However, studies that rely upon this simple binary characterization (89) to correlate methylation with expression find only modest negative correlations with expression levels (70, 93, 95).

The most common approach to associate DNA methylation and expression change is to first identify differentially methylated regions (DMRs) and then associate them with nearby genes. Numerous statistical tools have been developed to identify DMRs (89). Generally, DMRs are found by segmenting the genome into equally spaced regions and identifying which regions have statistically significant differences in methylation. DMRs are then associated with genes or other genomic regulatory elements within a certain distance to gain biological insight into their potential function. While DMR-based methods have been critically important in identifying imprinted loci (140), studies often find only weak correlations between DMRs near gene promoters and differential gene expression (93-95). One drawback of DMR methods is that they rely on a set of arbitrarily defined thresholds for the size and number of CpGs to include in the DMR. It is often recommended to adjust these parameters for each individual dataset since the

choice of these parameters has substantial implications in the numbers of DMRs identified and putatively associated genes.

One possible reason DMR methods fail to find a strong association between differential methylation and expression is they reduce DNA methylation to a single differential value removed from its local context. Recent work, however, has indicated that a large number of methylation patterns associate with differential gene expression (96). For example, methylation at CpG island-shores, regions of decreased CpG density flanking CpG islands, correlate with differential gene expression in colon cancer (69). Further, long hypomethylated domains in cancer often contain down-regulated genes (69). Positive correlations between gene body methylation and gene expression have also been frequently observed (141, 142).

Here, we present a new approach to predict gene expression changes that accounts for all methylation changes around the TSS. We have previously shown the importance of capturing methylation changes around the TSS to find patterns of methylation change that associate with expression changes using an unsupervised approach (96-98). We now build upon these results to develop a supervised method called ME-Class (Methylation-based Expression Classification), which classifies differential expression using signatures of differential methylation.

We use ME-Class to investigate alternate representations of DNA methylation and CpG density to identify methylation features that are most important in predicting expression change using data from the Roadmap Epigenomics Project. We then use ME-Class to examine the role methylation associated expression changes play in hematopoiesis using data from the Blueprint Epigenome project. Lastly, we demonstrate that ME-Class can identify a set of genes with

cancer-specific expression-associated DNA methylation changes that are silenced in tumor cells, but that are re-expressed when methylation is removed.

3.3 MATERIALS AND METHODS

3.3.1 Roadmap Epigenome Project (REP) WGBS and mRNA-seq

Samples from 17 primary tissues with matched whole genome bisulfite sequencing (WGBS) and RNA-seq were obtained from the Roadmap Epigenomics Project (REP, Supplementary Table 3.1) (1). Fractional methylation (M) is defined as mCG/CG. Differential methylation (ΔM) is defined as: $\Delta M = M_{S_2} - M_{S_1}$, where S_1 and S_2 correspond to the first and second sample in the differential comparison, respectively. We obtained consolidated methylation data, which was previously cross-assay standardized and uniformly processed. All CpG sites were filtered for 4x coverage or greater and analysis was performed using the hg19 genome assembly according to analysis standards established in REP. We used uniformly processed protein-coding gene level annotations from Genecode V10 to obtain standardized FPKM values. Each Genecode V10 annotation was converted to RefSeq annotations using the mygene python package (143). To create a standardized gene set with high quality methylation data, we excluded genes with ambiguous or incomplete TSS annotations, genes shorter than 5kb, genes with <40 CpGs assayed within +/-5kb of the TSS, genes where all CpGs within +/-5kb of the TSS had less than 0.2 methylation change, and alternative promoters. These filters were used to exclude non-coding and pseudogenes, genes shorter than the interpolation boundary, genes with low numbers of CpGs to reduce bias caused by individual CpGs, and genes with no methylation changes at their promoter, respectively. We only included RefSeq genes with cdsStartStat (n= 47637 genes)

and cdsEndStat (n= 47621 genes) with ‘cmpl’ according to the UCSC Table Browser. For any RefSeq genes with multiple RefSeq IDs corresponding to the same TSS location, we used a single RefSeq ID with the lowest accession number and excluded the remainder. This is a conservative method to simplify the annotations of genes with alternative promoter annotations. In analyses with the ROI classifier (see below), all genes with less than 4 exons were removed from analysis. Differentially expressed genes were defined as genes with ≥2-fold difference between samples after an applied floor of 5 FPKM to provide a conservative estimate of expression change. These filtering criteria have minimal effect on the fraction of CpG Island (CGI) -associated promoters, which went from 65.9% of genes before filtering to 67.3% after. CpG Islands were defined based on the CGI track from the UCSC Genome Browser (144). A full summary of filtered gene counts is in Supplementary Table 3.2.

3.3.2 Blueprint Epigenome project WGBS and mRNA-seq

WGBS and RNA-seq from 32 venous and cord blood samples were obtained from the Blueprint Epigenome project (2). Genome coordinates from hg38 were converted to hg19 using liftOver (144). All other analysis steps were performed identically to the REP data above.

3.3.3 Cancer WGBS and mRNA-seq

Breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), rectum adenocarcinoma (READ), and uterine corpus endometrial carcinoma (UCEC) matched normal-tumor WGBS and mRNA-seq samples were obtained from The Cancer Genome Atlas (TCGA) to train a model of tumorigenesis. Normal sigmoid colon (E106) WGBS and mRNA-seq data were obtained from

REP (1). HCT116 DKO1 (bi-allelic knockout of DNMT1 and DNMT3b) WGBS and mRNA-seq data were obtained from Blattler et al (145). All other analysis steps were performed identically to the REP data above.

3.3.4 Single Window (SW)

We computed the average methylation across a single, fixed ±1kb window around the TSS of each gene (69, 146). We performed logistic regression to predict differential expression from the average methylation change around the TSS (Figure 3.1a,b). Logistic regression cross-validation was run with 1000 maximum iterations of the optimization algorithm.

3.3.5 Differentially Methylated Regions (DMRs)

We used DSS-single to compare DMRs between individual samples (147). We identified DMRs ($p<0.01$) and used their size (bp), average differential methylation, and stranded distance (bp) to the closest TSS (0 if overlapping by >=1bp) as features for gene expression change classification with a Random Forest (RF) classifier with 1001 estimators (Figure 3.1a,b).

3.3.6 Regions of Interest (ROI)

The Regions of Interest (ROI) classifier reduces DNA methylation to multiple averaged values across annotated gene elements (upstream, exon, intron, and downstream bins) as features for a Random Forest (RF) to predict expression class (Figure 3.1c). ROI classifier features were implemented as described in Lou et al. (87) to predict differential expression class rather than

single sample binned expression values. We used a RF classifier with 100 trees as originally described. Increasing the number of trees to 1001 increased run time substantially without an appreciable increase in performance (Supplementary Figure 3.1).

3.3.7 ME-Class

Gene signatures were constructed as in VanderKraats et al. with minor modification (96). This signature allows the model to incorporate the entire profile of methylation changes across the gene's promoter including any CGI and CGI-shore regions (Figure 3.1a,d). In addition, these signatures allow comparison of methylation differences between genes, which have CpGs in different locations. We applied a localized z-score normalization of each differential methylation value in a 10kb window surrounding the TSS based upon the distribution of methylation values in a 100kb surrounding anchor window. We created methylation signatures using a piecewise cubic hermite interpolating polynomial (PCHIP) to interpolate a curve of z-score normalized differential methylation values in the 10kb window around the TSS for each differentially expressed gene. The interpolated curve was then subjected to Gaussian smoothing with a bandwidth of 50bp. Since CpG methylation values are highly autocorrelated (29), interpolation and smoothing of the data decrease the influence of sequencing error at individual CpGs (96). Similar smoothing approaches have shown a marked improvement in the ability to determine DMRs (148). To obtain discrete features, we subsampled our interpolated methylation signature at 20bp resolution. We then used these features with a RF classifier with 1001 estimators. We initially chose a RF classifier since it provides a nonparametric model, has a low number of hyperparameters, generates an internal unbiased estimate of testing error, identifies feature importance, and typically performs near-optimally with minimal tuning (149). Logistic

Regression (LR) (max_iter = 1001), Gradient Boosted Classification Trees (GBCT) (n_estimators=1001), Gaussian Naïve Bayes (NB), L2 distance-based k-Nearest Neighbors (kNN) (k=21), and Dynamic Time Warping (DTW) kNN (k=21) were implemented with default parameters other than stated modifications. All machine learning methods were implemented with scikit-learn and mlpy (DTW only) python packages (150, 151).

3.3.8 Whole Gene Methylation Models

We also implemented three alternative representations of methylation data to incorporate the full profile of methylation changes across the entire gene (Supplementary Figure 3.2a): Whole Scaled Gene (WSG), Whole Gene (WG), and Uniform Gene Features (UGF). For each representation, we created 125 bins in the regions 5kb upstream of the TSS and downstream of the TES (Transcription End Site) (20bp resolution). These regions/features were then added to specific features for each representation as follows. The WSG representation is an emerging representation in the literature to describe methylation changes by linearly scaling the methylation profile across the entire gene (1, 145, 152). To obtain discrete features, we used 500 bins across the gene (Supplementary Figure 3.2b). For the WG representation, we modeled methylation data as a curve (subsampled to 20bp resolution) across the entire length of the gene (Supplementary Figure 3.2c). For the UGF approach, we represented each exon with 10 scaled bins and each intron with 30 scaled bins. Multiple exons or introns were not averaged together (Supplementary Figure 3.2d). For WSG, we used a RF classifier. For both WG and UGF, we used curve similarity as defined with DTW (153), and classified expression changes using kNN (k=21).

3.3.9 Classifier Performance

To evaluate the amount of data needed to train ME-Class, we divided the 17 REP datasets into eight samples that were held out for evaluation and nine samples that were used for training. For a given number of training samples (*n*), nine random permutations of n pairwise comparisons were chosen from the training samples and used to train nine ME-Class classifiers. The resulting nine classifiers were then evaluated on a fixed set of eight comparisons from the holdout evaluation samples (Supplementary Figure 3.10).

To evaluate each classifier, we implemented a conservative two-stage cross-validation framework (Figure 3.1e) to ensure that the model does not overfit to any given sample or individual gene. For a given evaluation, we performed the following procedure: 1) Leave-one-out sample pair cross validation: We divided all differential training samples into a training set and an evaluation set. This ensured that no individual sample from the training set appears in the evaluation set. 2) 10-fold gene cross validation: We randomly divided the genes from the evaluation sample into 10-folds. To evaluate each fold, we first removed examples of all genes in the evaluation fold from all samples in the training set prior to training. Thus, if gene A is in the evaluation data, no examples of gene A for any tissue are used for training. We then trained on the training samples/genes and evaluated the chosen fold of genes in the evaluation samples. We then repeated this process 10 times for each fold of the evaluation sample and for each differential sample in the dataset. This process helped the classifier generalize across genes and samples by ensuring that each evaluation gene does not observe either an example of itself or any other genes from its individual sample.

Average performance based on the accuracy, reject rate, positive predictive value (PPV), and negative predictive value (NPV) was reported across all genes treated as a single pool from all

samples. Testing accuracy was defined as the number of genes with correctly predicted expression divided by total genes returned. Receiver Operator Characteristic (ROC) and Precision Recall (PR) curves were averaged to provide a sample-wise level of reporting. RF feature importance was estimated as Gini importance. Unless otherwise stated, all statistical comparisons were performed using FDR-corrected paired, pairwise Wilcoxon rank sum test in R.

3.3.10 Gene Ontology Analysis

Gene Ontology analysis was performed by analyzing gene lists with Functional Annotation Clustering with default parameters from DAVID (154). Blueprint analysis gene lists were identified by any genes with ≥90% probability of classification in ≥2 differential samples. Colon cancer gene lists were identified by any down-regulated genes with a ≥90% probability of classification by ME-Class (see discussion of reject rate in Results).

3.4 RESULTS

3.4.1 ME-Class predicts gene expression change from differential methylation in tissue samples

Since the goal of most genome-wide methylation studies is to identify how changes in methylation alter expression, we examined the ability of methylation to predict differential expression change. We first sought to understand whether a methylation signature approach (i.e. modeling the entirety of methylation changes around a gene's TSS) could outperform current DMR, single window, and region of interest (ROI) methods in finding genes with associated differential methylation and expression. (Figure 3.1, Table 3.1).

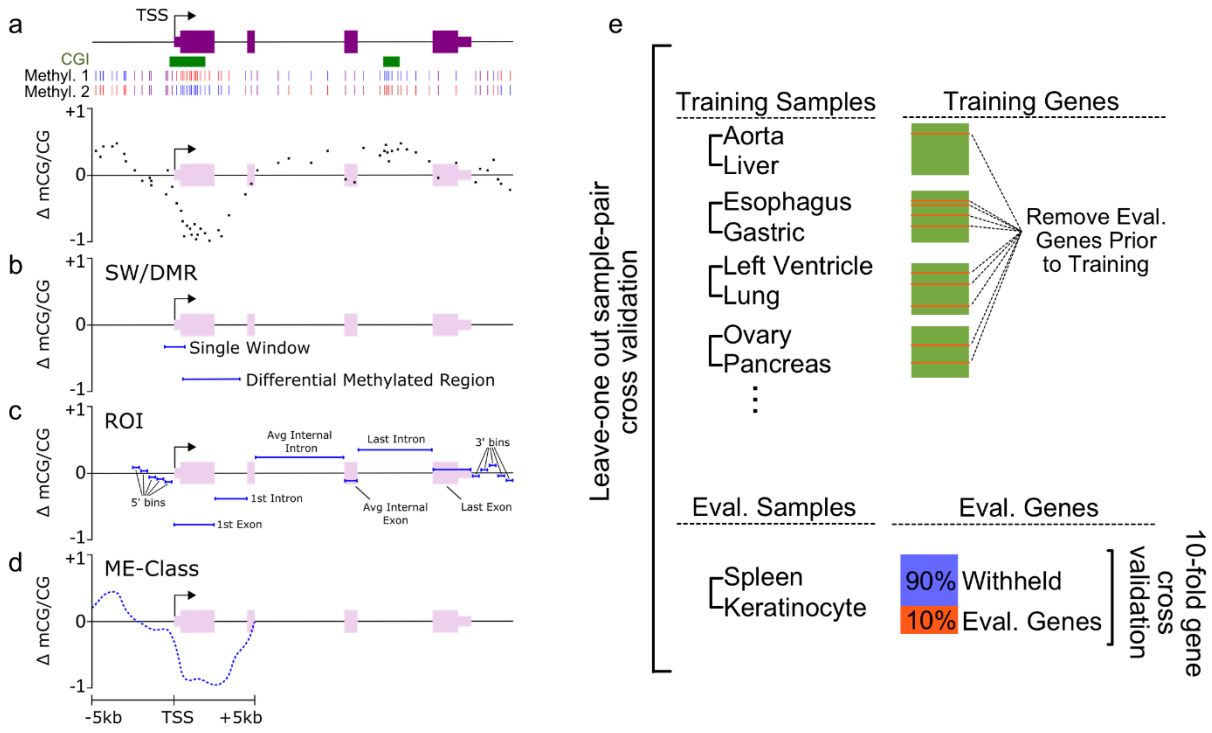


Figure 3.1 Models of DNA methylation and validation framework for predicting differential gene expression change from differential DNA methylation. a) Heat map indicates methylation status at individual CpG sites – red is fully methylated, blue is fully unmethylated – for an example gene in two samples (Methyl. 1 and Methyl. 2). Individual points below indicate differential DNA methylation (Methyl. 2 – Methyl. 1) across the example gene at individual CpG sites. b) Example regions that would be used to calculate the single window (SW) and differentially methylated region (DMR) using the data in (a). c) Regions used to calculate methylation features for the Region of Interest (ROI) representation of the gene in (a). d) ME-Class representation of the gene in (a). Each individual point is the differential methylation value used as a feature in a Random Forest after interpolation and smoothing. e) Cross-validation comparison framework. Evaluation is performed sample-wise across the 17 sample comparisons. In the evaluation comparison, genes are split into 10-folds. Prior to training, each evaluation gene is removed for all other tissues. Further model details are in Table 3.1. CGI = CpG island.

Table 3.1 DNA methylation features and classification methods for each model in Figure 3.1.

DNA Methylation Representation	Features	Classification Method
Single Window (SW)	Δ mCG/CG +/-1kb of TSS	Logistic Regression (LR)
Differentially Methylated Regions (DMR)	Distance from TSS to DMR (bp) DMR width (bp) Avg. Δ mCG/CG	Random Forest (RF)
Regions of Interest (ROI)	Avg. Δ mCG/CG: Five 400bp bins 5' of TSS 1st Exon 1st Intron Avg. Internal Exon Avg. Internal Intron Last Exon Last Intron Five 400bp bins 3' of txEnd	Random Forest (RF)
Methylation-based Expression Classification (ME-Class)	Δ mCG/CG of 500 bins (20bp) +/-5kb of TSS	Random Forest (RF)

To compare supervised classifiers, we used whole genome bisulfite sequencing (WGBS) DNA methylation and RNA-seq data from the Roadmap Epigenomics Project for 17 tissue samples (Supplementary Table 3.1) (1). We implemented a conservative sample-wise and 10-fold gene-wise cross-validation framework that ensures the genes in the evaluation step have not been seen in any tissue during training (Figure 3.1e). Since patterns of differential DNA methylation can be very similar between datasets, this evaluation framework tests the strength of the DNA

methylation representation and the universality of DNA methylation patterns rather than the ability to simply recall an observed gene's methylation signature.

Using this framework, ME-Class outperformed all methods by receiver operator characteristic (ROC) curve analysis ($p < 10^{-3}$ for ME-Class compared to each of DMR, SW, and ROI, Figure 3.2a) and precision-recall (PR) analysis ($p < 10^{-3}$ for ME-Class compared to each of DMR, SW, and ROI, Supplementary Figure 3.3c). In addition, ME-Class performed better than or equal to each of the other methods analyzed for any individual comparison based on ROC AUC (area under the curve; Figure 3.2c). Interestingly, there was a large variability in the classification performance of the differential samples. While in the case of keratinocyte comparisons this could likely be explained by poor sequencing coverage (Figure 3.2e,f), this per-sample performance difference appeared primarily to be due to fundamental differences in the methylation profiles of the expression classes. The ROC AUC was strongly correlated (Figure 3.2d, $R^2=0.87$) with the average methylation difference between up- and down-regulated genes in the region +0.5kb to +2.5kb relative to the TSS (the most important region for classification, Figure 3.2b). While we cannot rule out there is some other technical artifact in the data causing this correlation, it appears that most normal tissue methylation-based expression classification derives from TSS 3' proximal methylation changes.

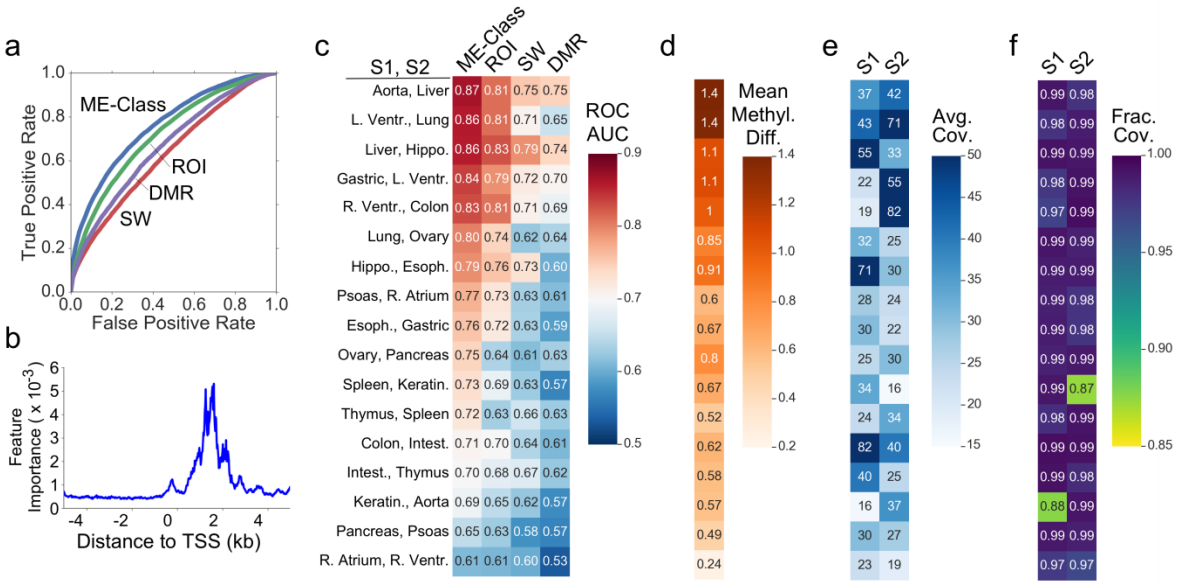


Figure 3.2 ME-Class outperforms standard methods for tissue-specific expression classification. Methods were evaluated using 17 tissue samples from the REP with the two-stage cross-validation framework in Figure 3.1e. a) ROC analysis from a combination of all 17 samples (ROC AUC: ME-Class, 0.76; ROI, 0.71; DMR, 0.63; SW, 0.66). b) RF feature importance from ME-Class trained on all 17 differential comparisons from REP. c) ROC AUC for each of the 17 samples comparisons. d) Mean z-score normalized methylation difference of the region [+0.5kb, +2.5kb] relative to the TSS. e) Average CpG coverage and f) average fraction of CpGs within the 10kb window around the TSS for each REP sample. S1 and S2 correspond to the first and second sample in the evaluation differential comparison. L. Ventr. = Left ventricle, Hippo. = Hippocampus, R. Ventr. = Right ventricle, Esoph. = Esophagus, R. Atrium = Right Atrium, Keratin. = Keratinocyte, Intest. = Intestine.

3.4.2 ME-Class generates a list of genes with associated differential methylation and expression

Transcription can be influenced by multiple factors other than DNA methylation, such as transcription factors or chromatin modifications. Thus, the key issue is whether ME-Class can be used to identify a subset of genes that have high quality associations between differential methylation and expression. For this purpose, we introduced a reject rate into the classifier that allows us to control for external factors other than methylation that indicate gene expression. The reject rate excluded genes that cannot be reliably predicted (i.e. they likely do not have methylation-associated expression changes) using a threshold for the probability of classification output by each classifier. In practical terms, the reject rate allows one to set a parameter based on the cross-validation evaluation error that can control the false positive rate when running ME-Class on unseen samples. In Figure 3.3a, we observe that ME-Class outperformed ROI, SW, and DMR methods in accuracy and proportion of the genes returned across all rejection rates.

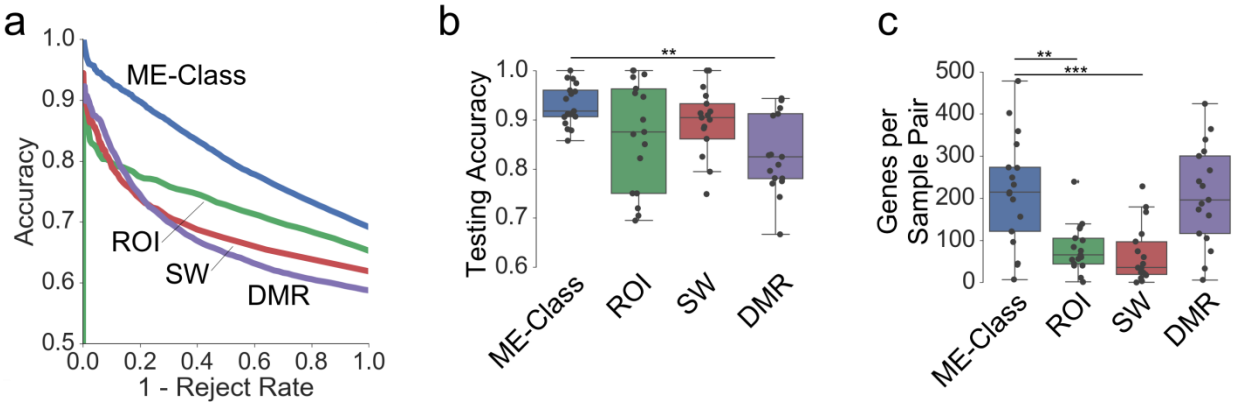


Figure 3.3 ME-Class identifies more genes at higher accuracy with expression-associated methylation changes in tissue-specific differential comparisons in the REP dataset. a) Classifier

accuracy versus 1 - reject rate. b) Accuracy of testing sample at 90% operating probability of classification. c) Number of genes identified with expression-associated methylation changes at 90% operating probability of classification. Points in (b) and (c) indicate the performance of individual REP sample comparisons. ** indicates $p < 0.005$ and *** indicates $p < 0.001$. All other comparisons were not significant for $\alpha=0.05$.

We next set the classification probability at 90% and examined how many genes were returned by each method and the accuracy of this list. ME-Class returned significantly more genes than the SW and ROI methods (ROI: $p=2.9 \times 10^{-3}$; SW: $p=9.2 \times 10^{-4}$) and was significantly more accurate at 90% probability of classification than the DMR method (DMR: $p=3.0 \times 10^{-3}$) (Figure 3.3b,c). ME-Class returned the largest average number of genes, 217, at the highest level of accuracy (93.1%). The ROI and SW methods achieved lower levels of accuracy and returned a much lower average number of genes (ROI: 81 genes, 86.9% accuracy; SW: 66 genes, 84.5% accuracy)**.** The DMR method returned a similar average number of genes, 207, but at the cost of a much lower level of accuracy (83.3%). This implies that the nearest isolated DMR is often insufficient to predict the expression class even when tuning the DMR parameters to find optimal segmentation parameters.

At 90% probability of classification, ME-Class did not show any bias towards the positive (up-regulated) or negative (down-regulated) class (Supplementary Figure 3.3a,b). ME-Class matched or exceeded the accuracy given the probability of classification, indicating that this probability can be used as an estimate of the final classification error in cross-sample comparisons (Supplementary Figure 3.3d). This demonstrates that when running ME-Class on new samples,

the probability of classification can be used as an estimate of the false discovery rate. Meanwhile, the ROI, SW, and DMR approaches all have a lower accuracy than the probability of classification at high probabilities indicating that they are likely overfit and do not generalize well to other datasets.

To understand the difference in why ME-Class was highly predictive of some genes and not others, we examined metagene plots of the methylation signal at predicted genes subset by the probability of prediction. We observed that the highest predicted genes have the greatest methylation difference between downregulated and upregulated genes in a [+0.5kb, +2.5kb] region around the TSS of each gene, and that this signal decays with decreasing probability (Supplementary Figure 3.4). This dampening of the methylation signal can likely be attributed to either noise from the WGBS assay or biological noise from cell-type heterogeneity.

3.4.3 3' proximal and TSS regions are most predictive of differential expression

We next sought to understand why ME-Class performed better by examining which features are most important for classification. We first developed a series of ME-Class classifiers each using signatures from 5kb windows centered at varying distances away from the TSS (Figure 3.4a,b). Performance peaks for a window centered 2kb downstream of the TSS, indicating that the most important features exist downstream of the TSS. There was no substantial difference observed in ROC AUC between the entire 10kb window compared for the 2kb downstream centered 5kb window (Figure 3.4b). This agrees with our analysis of RF feature importance, which showed that the most important features for gene expression classification occur downstream +0.5 to +2.5kb of the TSS (Figure 3.2b).

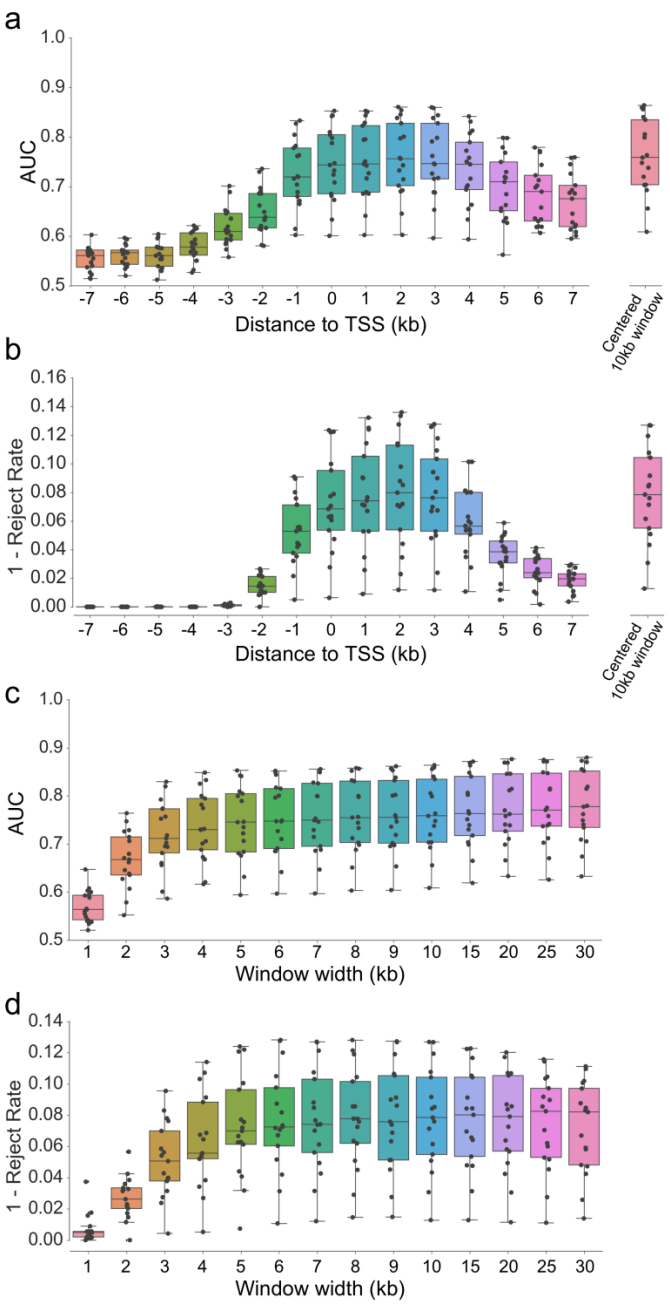


Figure 3.4 Importance of DNA methylation changes 3' proximal to TSS for tissue-specific expression classification. Methods were evaluated using 17 REP tissue samples with the two-stage cross-validation framework in Figure 3.1e. a) ROC AUC and b) 1-reject rate for ME-Class

methylation signatures created from fixed 5 kb windows centered at varying distances to the TSS. c) ROC AUC and d) 1-reject rate for ME-Class methylation signatures created using increasing window widths centered at the TSS. Individual points are the performance of individual REP sample comparisons.

We next evaluated the use of a Most Important (MI) window within the [+0.5kb, 2.5kb] region around the TSS of each gene in the REP dataset (Supplementary Figure 3.5). Both a SW classifier trained on only this region, and a ME-Class classifier using only features from this region performed substantially worse than ME-Class given data from the full [-5kb, +5kb] region (Supplementary Figure 3.5), underscoring the importance of using all the data around the TSS to represent DNA methylation.

We then designed a series of ME-Class classifiers to examine how the size of the TSS-centered window affects performance. Increasing the window size beyond 10kb did not show substantial improvements in ROC AUC (Figure 3.4c). However, after the window size increases greater than 10 kb there is a decrease in the number of genes returned at 90% probability of classification ($p=4.5 \times 10^{-3}$) (Figure 3.4d). Combined, these results indicate that high-resolution features across a window of at least 5kb wide and shifted 3' proximal of the TSS is sufficient to capture the complexity of methylation signal around the promoter required for expression prediction in the REP samples.

3.4.4 Alternative models and features to improve ME-Class

We next sought to determine whether the underlying model used in ME-Class was sufficient for predicting expression. We were inspired from our previous unsupervised analysis (96) to design ME-Class to model the changes in methylation levels around the TSS. However, other features, including CpG density, the density of methylated CpG sites, and gene body methylation, have been described in the literature to have correlations with expression. Thus, we sought to understand whether adding these features would improve classification performance.

3.4.5 CpG density does not improve ME-Class but CpG-poor genes are more predictive

The density of methylated CpGs has been implied as the important feature for why CGI methylation affects gene silencing (155, 156). We thus compared gene signatures computed from the normalized methylation density (mCG/bp), CpG density, or fractional methylation (mCG/CG) to see which feature performed best. ROC and PR analysis, as well as examining the relationship between accuracy and reject rate, show no substantial increased effect of using mCG/CG rather than mCG/bp. Unsurprisingly, a model based on CpG density alone performs nearly equivalent to random guessing (Supplementary Figure 3.6a). While the direct addition of CpG density did not improve performance, we found that ME-Class performed worse on CGI-associated and CpG-rich promoters (Supplementary Figure 3.7). This is in agreement with prior findings that there is a stronger correlation between methylation and expression for genes which had no CpG island as compared to those with CpG Islands (85). However, since more genes contain CGIs, ME-Class identifies a strong association (90% probability of prediction) between differential methylation and expression for more CGI-associated genes (mean=135 for REP samples) than CGI-poor genes (mean=108). A similar trend is observed for CpG-rich

(mean=170) versus CpG-poor (mean=73) promoters. Previously, it has been hypothesized that CGI-associated and CpG-rich genes tend to remain unmethylated in normal cell-types irrespective of their expression levels (56); however, our analysis suggests that while there are better associations between methylation and expression for non-CGI-associated genes, there are more CGI-associated genes that show strong associations.

3.4.6 The addition of gene body methylation changes does not improve ME-Class

Since gene body methylation has been shown to be positively correlated with gene expression (85, 141, 142), we examined if we could improve ME-Class by adding additional gene features that modeled methylation changes in the gene body. ME-Class performance was not substantially improved by adding features for averaged gene features similar to that of the ROI method such as the average methylation of internal exons, introns, and region downstream of the gene (Supplementary Figure 3.6b). This was unsurprising, since feature importance analysis of the ROI classifier indicated that the most important features for classification were the methylation levels of the first exon and first intron, which substantially overlap the region from the TSS to +5kb.

We also investigated whether other gene representations could determine whether methylation information from the gene body could improve classification performance. Therefore, we implemented three alternative approaches to model DNA methylation throughout a gene: Whole Scaled Gene (WSG), Whole Gene (WG), and Uniform Gene Features (UGF) (Supplementary Figure 3.2b-d). In the Whole Scaled Gene (WSG) approach, the methylation profile is interpolated across the entire gene and then all genes are rescaled to a uniform length. This is a common method to visualize genomic trends in genome-wide methylation data (1, 145, 152). WG is similar, but the genes are not scaled after interpolation. Lastly, in the UGF approach

methylation is interpolated then methylation features are extracted using a uniform number of bins for each exon and intron. All alternative approaches include regions -5kb of the TSS and +5kb of the TES. Using ROC AUC analysis, the TSS-centric (default ME-Class representation) model outperforms the WG ($p=3.7 \times 10^{-5}$) and UGF ($p=2.3 \times 10^{-5}$) approaches (Supplementary Figure 3.2e-g). Further, the TSS-centric approach identifies more genes on average (TSS: 178, WSG: 112) at a higher average level of accuracy (TSS: 93%, WSG: 89%) than the WSG approach.

All combined, these results suggest that models that incorporate gene body methylation, whether through average features or whole gene representations, do not substantially outperform models comprising only information from around the TSS. Our results demonstrate that features in the gene body and downstream of the TES are minimally important when using differential methylation to classify expression change. Thus, even though there are correlations between differential gene body methylation and differential expression, there is minimal new information in the gene body relative to the information already found in the +/-5kb region around the TSS.

3.4.7 Optimizing ME-Class

Before using ME-Class to analyze additional samples, we examined whether we could tune ME-class for better performance. Using the REP dataset, we compared the performance of the RF, Logistic Regression, Gradient Boosted Classification Trees (GBCT), Naïve Bayes, and k-NN. We also compared the RF-based approach to a method using DTW as a curve similarity metric for kNN classification. RF, Logistic Regression and GBCT outperformed the remaining machine learning methods by both ROC AUC analysis and examining the relationship between accuracy

and reject rate (Supplementary Figure 3.8). We also found that ME-Class performed similarly well as long as smoothing parameters were maintained below 200 bp (Supplementary Figure 3.9a,b). Interpolation and smoothing serve to decrease inaccuracy of low coverage methylation calls, as has been observed in DMR callers (148). Changes in the interpolation method also had no substantial effect on performance (Supplementary Figure 3.9c). To assess how much training data ME-Class requires for accurate classification, we separated the REP dataset into 9 differential samples for training held out 8 differential samples for evaluation (see details in Materials and Methods). ME-Class was consistent within 0.02 ROC AUC of the full training set after using 3 samples (Supplementary Figure 3.10a) and showed minimal increases in obtaining consistent gene sets as the full training set when using more than 5 samples (Supplementary Figure 3.10b) or 20,000 genes (Supplementary Figure 3.10c). We also observed that the RF outperforms the DTW in identifying consistent gene sets as the full training set when using more than 2 samples (Supplementary Figure 3.10d).

3.4.8 Myeloid/Lymphoid differential methylation comparisons are most predictive of expression change

We next applied ME-Class to identify methylation-associated expression changes in hematopoiesis. We used WGBS and mRNA-seq datasets provided by the Blueprint Epigenome project in cord and venous blood composed of 32 isolated samples from 10 cell types. We retrained ME-Class using the entire 17 samples from the REP data above and then used this model to find methylation-associated expression changes in 469 hematopoietic lineage-wise differential comparisons. We found a large variation in the number of genes identified based on the cell-types being compared. Comparison of relatively distantly related lymphoid and myeloid

lineages resulted in 54-218 genes (mean = 88) returned at 90% probability of classification (Figure 3.5a,b). Gene ontology analysis suggests that genes identified in myeloid-lymphoid comparisons are enriched for genes involved in T-cell activation, leukocyte differentiation and hematopoiesis. Similar performance results were found if we characterize the classifier based on ROC AUC (Figure 3.5b). This contrasts with a comparison of closely related myeloid cells such as macrophages, monocytes and dendritic cells, which identified between 0 and 55 genes when comparing any two cell types (mean = 16 genes; $p=2 \times 10^{-16}$) and demonstrate performance near random guessing (ROC AUC = 0.55±0.04). Neutrophil samples stood out as particularly poor performers in all comparisons suggesting that either there are not methylation-associated expression differences in these cells, or that methylation profiles in these cells are fundamentally different from that of other tissues (Supplementary Figure 3.11).

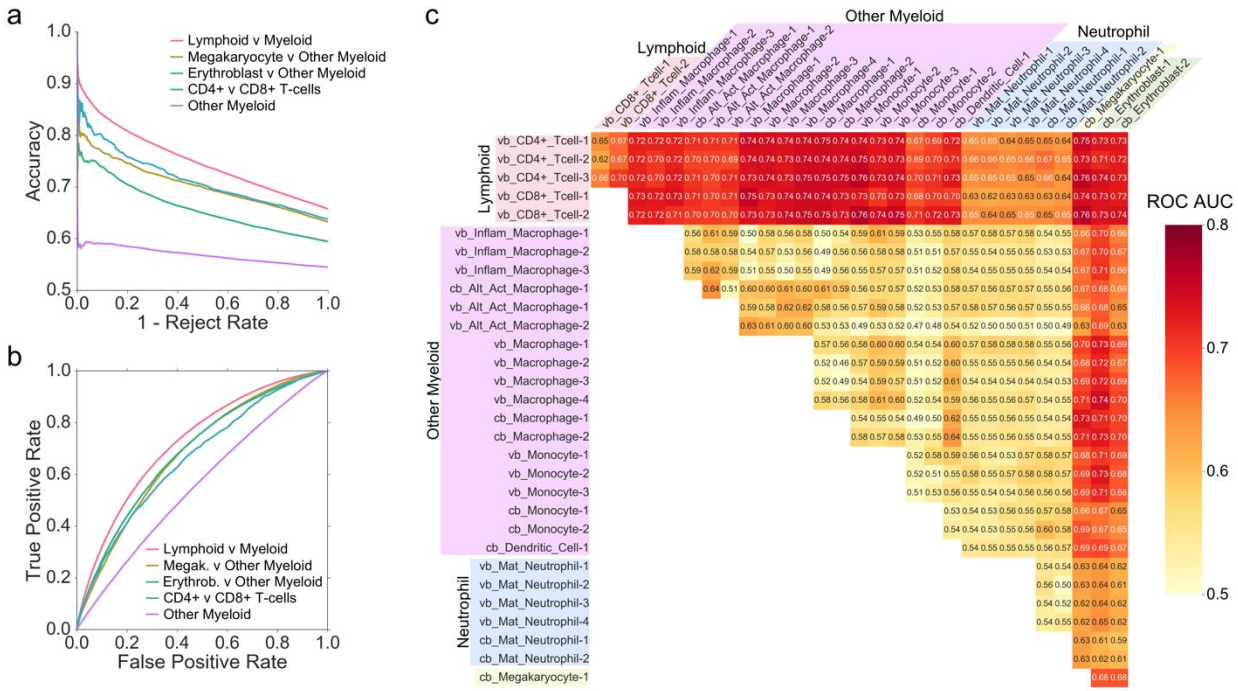


Figure 3.5 ME-Class performance is higher for cell comparisons between distally related cell lineages as opposed to directly related ones. a) Accuracy versus 1- reject rate and b) ROC of selected cell-types (ROC AUC: Lymphoid v Myeloid, 0.72±0.02; Megak. v Other Myeloid, 0.68±0.02; Erythrob. v Other Myeloid, 0.68±0.02; CD4+ v CD8+ T cells, 0.66±0.02; Other Myeloid, 0.55±0.03). c) ME-Class ROC AUC performance for each sample-wise comparison of hematopoiesis samples. Vb = venous blood, cb = cord blood, Erythrob. = Erythroblast, Alt_Act_Macrophage = Alternating activated Macrophage, Inflam_Macrophage = Inflammatory Macrophage, Mat_Neutrophil = Mature Neutrophil, Megak. = Megakaryocyte. ROC AUC error is the standard deviation.

Based on ROC analysis there was an inverse relationship between the relatedness of the cells being compared and the ROC AUC from ME-Class (Figure 3.5c). Similar results were also obtained using an ME-class model trained a combination of closely- and distantly-related hematopoietic cell types (Supplementary Figure 3.12). While other analyses have examined associations upon myeloid (157) and B-cell differentiation (158), this analysis demonstrates that truly predictive differences in differential DNA methylation primarily reside between the myeloid and lymphoid lineages, the two major lineages derived from hematopoietic stem cells.

3.4.9 ME-Class identifies subsets of genes sensitive to demethylation in colon cancer

We next examined whether ME-Class can accurately identify genes that are hypermethylated and silenced in a model of cancer. For this problem, we first trained a cancer-specific ME-Class model using WGBS and mRNA-seq data from four different normal-tumor differential

comparisons from The Cancer Genome Atlas (TCGA) (BRCA, COAD, READ, and UCEC). We then used this cancer-specific ME-Class model to identify methylation-associated changes in expression in a colon cancer cell line (HCT116) relative to normal colon tissue (Figure 3.6a).

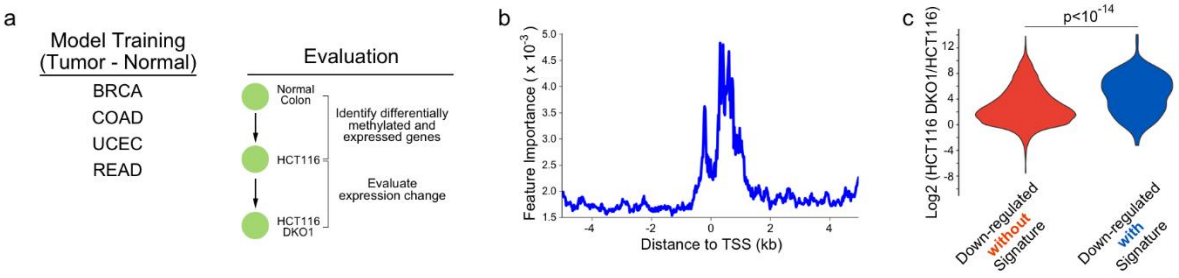


Figure 3.6 ME-Class identifies genes re-expressed after removal of DNA methylation in a model of colon cancer. a) Experimental scheme. We first built an ME-class model using WGBS data from four tumor-normal pairs from TCGA. We then used this model to identify genes with expression-associated methylation changes upon tumorigenesis (HCT116) from normal colon (REP) and evaluate the demethylation effect by genetic manipulation (DKO1: bi-allelic knockout of DNMT1 and DNMT3b in HCT116). b) Feature importance for the ME-Class training model built from four TCGA tumor-normal samples (COAD, BRCA, READ, UCEC). c) Violin plots showing the expression fold change in HCT116 DKO1 cells using down regulated differentially expressed gene sets identified with (n=187 genes) and without (n=5,370 genes) identified methylation signatures.

We observed that the primary peak of feature importance in our trained model shifts from the region [+500bp, +2500bp] downstream of the TSS in the REP tissue-specific model to the region

[-500bp, +1500bp] overlapping the TSS in the TCGA normal-tumor model (Figure 3.6b). We found a severe class imbalance; 187 genes with methylation-associated expression changes were identified as down-regulated, but no genes were predicted as up-regulated. Functional annotation clustering of gene ontology of the 187 down-regulated, identified genes showed that these genes are enriched for C2H2 zinc fingers, previously shown to be hypermethylated and silenced in carcinogenesis, and are involved in cell adhesion, whose dysregulation is important for tumorigenesis (159). To understand whether the tumor-associated hypermethylation was functional, we examined what happened to these genes after removal of methylation by double knockout of DNMT1 and 3b (HCT116 DKO1). Genes with an ME-Class signature showed a significant upregulation of expression relative to down-regulated genes not identified by ME-Class (Figure 3.6c, $p=1 \times 10^{-14}$). These results demonstrate that ME-Class can identify genes likely regulated by DNA methylation in human disease and that hypermethylation near the TSS plays a primary role in modulating gene activity in a model of colon cancer (56, 160, 161).

3.5 DISCUSSION

One challenge in the field of DNA methylation analysis has been the difficulty of integrative analysis of genome-wide DNA methylation and expression data. While methods exist to facilitate this task, they have many parameters that must be set, often with no clear way to make intelligent choices for their values. For example, DMR and SW approaches require the user to set several parameters (such as minimum window size or a minimum number of CpGs) that can drastically change the list of genes with predicted methylation changes. Our results from REP analysis show that these methods cannot be used to accurately predict expression from DNA methylation, likely because they cannot model the signal complexity necessary to associate

methylation and expression change (Figure 3.2). Incorporating the complexity of patterns, rather than reducing methylation to a single or even multiple averaged values, is critical for the success of ME-Class. Alternative gene representations that incorporate gene body methylation perform no better than representations that focus on the region around a gene's TSS (Supplementary Figure 3.6b). Thus, the information obtained from correlations between gene body methylation and expression is either too noisy, or it is redundant with information in the promoter region. In the future, it may be possible to improve ME-Class by adding features specific to enhancers, but first we need better computational tools and experimental data across multiple cell types to connect regulatory units with specific genes.

In this study, we asked whether we could build models of DNA methylation that would generalize across different genes and across samples. For this purpose, we established a strict evaluation framework to test changes in methylation to identify these most likely affected genes. Using a training and evaluation paradigm that consists of both cross-sample and cross-gene evaluation, we have shown that ME-Class predictions are not overfit to any given dataset. ME-Class performs well across high quality datasets without the need for tuning for individual datasets. However, there are limits to this generalization. For example, tissue-specific and cancer-specific datasets require different models to achieve high performance. Further, even though we excluded alternative promoters from this analysis to compare methods across a set of high quality reference genes, ME-Class can use isoform-specific promoter locations and expression data as input to provide an isoform-level analysis.

The role of gene-specific DNA methylation changes in development has been debated for many years (5). If promoter DNA methylation played a primary role in regulating cell-type specific expression changes then one would expect to observe a large fraction of differentially expressed

genes with methylation-associated expression changes even amongst closely related cell types. However, we do not observe this. On the other hand, if methylation were a consequence of expression change, then one would expect to see a large number of methylation-associated expression differences in distantly related cell types or tissues, but not in closely related ones. Our results from ME-class are consistent with the latter statement. ME-Class identifies on average 7.5% (217 genes/sample) of differentially expressed genes from different tissues in REP are associated with promoter methylation changes at 90% accuracy, and 2.5% (88 genes/sample) of differentially expressed genes for distantly related hematopoietic lineages (myeloid vs lymphoid). However, ME-class performs poorly on sample comparisons of closely related hematopoietic lineages, often identifying few genes (myeloid vs myeloid; mean = 0.6% or 16 genes/sample). Our data is thus consistent with a model where transcriptional changes precede methylation changes at the promoter during differentiation, and thus, tissue- and cell- specific DNA methylation changes are likely a consequence of transcriptional changes. While DNA methylation is unlikely to initiate tissue-specific expression, we cannot rule out the possibility that these methylation changes at the promoter play a later role in maintaining these transcriptional programs.

In contrast, the reactivation of genes identified by ME-Class as methylated and silenced upon removal of DNA methylation is consistent with the hypothesis that methylation changes at the promoter in cancer play a direct role in gene regulation. This observation is consistent with our recent work showing that an unsupervised analysis of differential methylation data in AML could identify a set of genes that were likely to be up-regulated upon treatment with demethylating agents (97). We also observe a shift in the most informative region for expression associated methylation changes from [+500bp,2.5kbp] in a tissue-specific model to [-500bp,

+1.5kbp] in a cancer-specific model. While early studies suggested that tissue-specific and cancer-specific expression-associated methylation changes were similar, our results are in agreement with more recent studies that use higher resolution methods and larger numbers of samples (85, 95, 162, 163). Further, the observation that both windows are shifted downstream of the TSS is in agreement with recent studies that have suggested that the transcriptional activator p300 can bind downstream of the TSS at unmethylated CpG Islands to increase gene expression (85), and that decreases in methylation downstream of the TSS co-occur with increases in active H3K4me3 that also shift downstream of the TSS (163, 164). This difference in expression-associated methylation changes may explain why methylation appears to play a role in gene silencing in cancer, but not in development. In addition, the context dependent nature of these models has a profound effect on downstream applications indicating that different methylation models may need to be trained for different contexts (i.e. cancer-specific models need to be trained to understand expression-associated methylation changes in cancer). Once these models are trained, they are applicable across other similar datasets.

As more large-scale, genome-wide DNA methylation studies of the differences between matched normal and tumor samples become available, tools such as ME-class will prove invaluable to understand how specific methylation changes affect transcription. In addition, our results show that ME-Class is a powerful tool to identify genes that are silenced by methylation in disease and could be used to facilitate the identification of patients who may benefit from clinically-approved demethylating therapeutics (165).

3.6 AVAILABILITY

ME-Class is publicly available on Github at http://github.com/cschlosberg/me-class

3.7 FUNDING

This work was supported by the Siteman Cancer Center, U.S. Department of Defense Congressionally Directed Medical Research Program for Breast Cancer [W81XWH-11-1-0401], and the National Institutes of Health [NIGMS 5R01GM108811, NLM R21LM011199] (to J.R.E.), and National Institutes of Health T32 Genome Analysis Training Program [2T32HG000045-16] for pre-doctoral support to C.E.S.

3.8 ACKNOWLEDGEMENTS

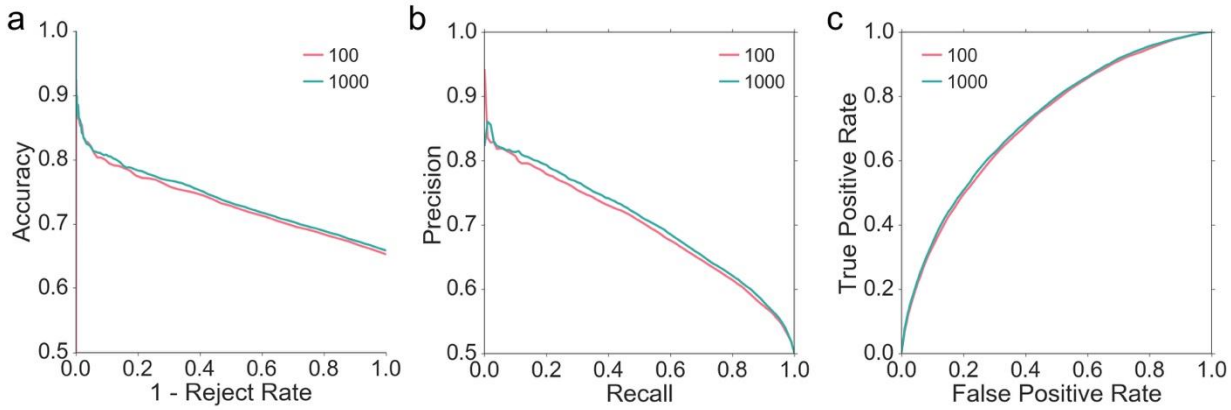
We would like to acknowledge Tao Ju and Kilian Weinberger for helpful discussion in the development of ME-Class. We further thank Jerry Fong, Lisa Rois, and Manoj Singh for assistance in testing the ME-Class code and critical feedback on this manuscript.

Supplementary Table 3.1 Roadmap Epigenomics Project (REP) biological and technical replicate counts.

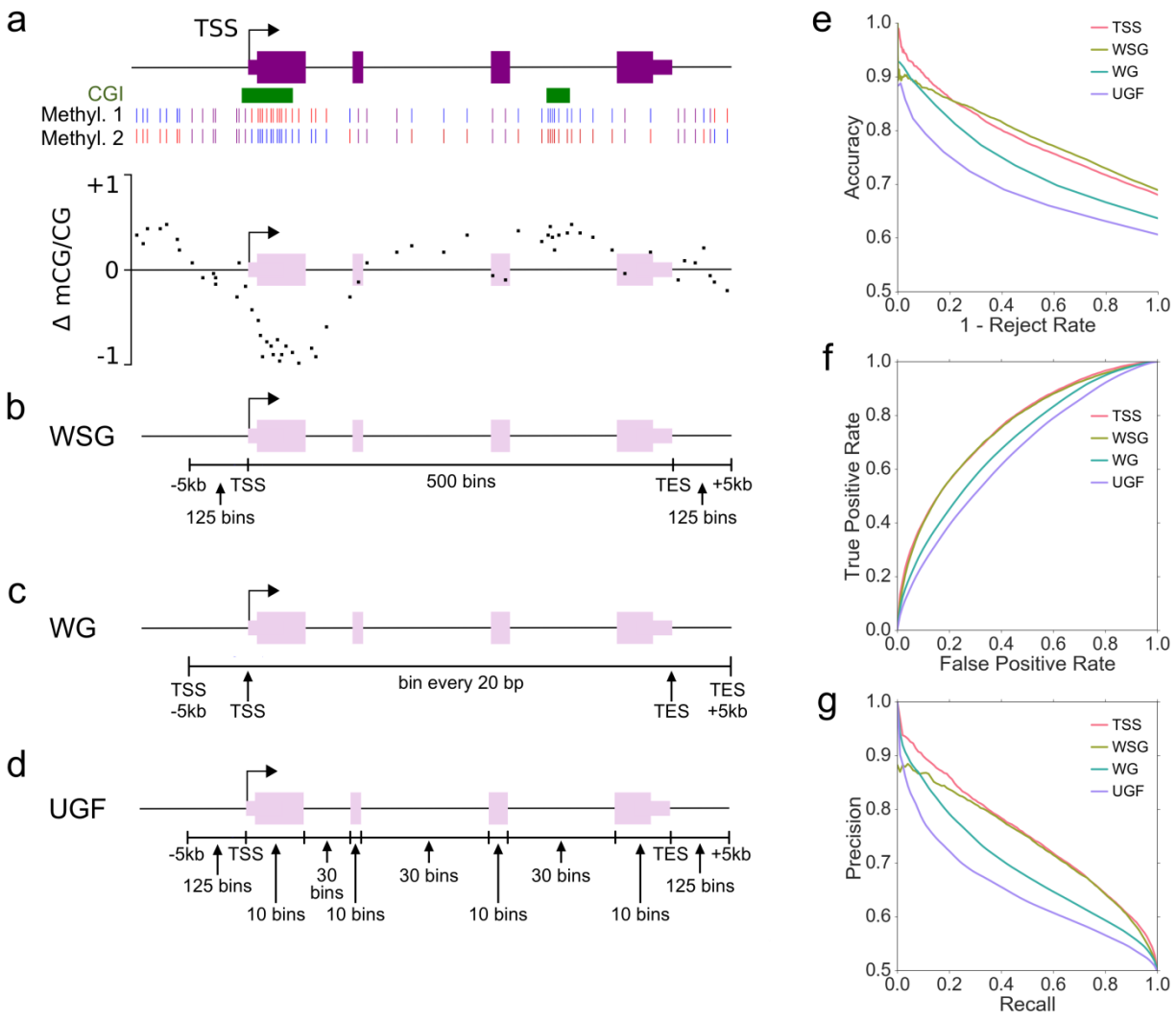
				Technical Replicates	
ID	**Tissue**	**Tissue Short Name**	**Sources**	**WGBS**	**RNA-seq**
E058	Penis Foreskin Keratinocyte	Keratin.	1	2	3
E065	Aorta	Aorta	1	6	2
E066	Adult Liver	Liver	3	2	2
E071	Brain Hippocampus Middle	Hippo.	2	3	2
E079	Esophagus	Esoph.	1	2	2
E094	Gastric	Gastric	1	5	3
E095	Left Ventricle	L. Ventr.	1	4	2
E096	Lung	Lung	1	2	2
E097	Ovary	Ovary	1	2	1
E098	Pancreas	Pancreas	1	2	2
E100	Psoas Muscle	Psoas	2	3	3
E104	Right Atrium	R. Atrium	1	3	1
E105	Right Ventricle	R. Ventr.	2	5	2
E106	Sigmoid Colon	Colon	2	2	3
E109	Small Intestine	Intest.	2	4	3
E112	Thymus	Thymus	1	2	1
E113	Spleen	Spleen	1	3	3

Supplementary Table 3.2 Differentially expressed and ME-Class interpolated gene counts from Roadmap Epigenomics Project (REP).

Tissue 1	Tissue 2	Diff. Expr. Genes	Interpolated Genes
Penis_Foreskin_Keratinocyte (E058)	Aorta (E065)	4596	3804
Aorta (E065)	Adult_Liver (E066)	3891	3186
Adult_Liver (E066)	Brain_Hippocampus_Middle (E071)	5020	4136
Brain_Hippocampus_Middle (E071)	Esophagus (E079)	4270	3484
Esophagus (E079)	Gastric (E094)	2801	2220
Gastric (E094)	Left_Ventricle (E095)	3438	2738
Left_Ventricle (E095)	Lung (E096)	3711	2942
Lung (E096)	Ovary (E097)	3227	2521
Ovary (E097)	Pancreas (E098)	3931	3148
Pancreas (E098)	Psoas_Muscle (E100)	5240	4285
Psoas_Muscle (E100)	Right_Atrium (E104)	3437	2854
Right_Atrium (E104)	Right_Ventricle (E105)	1128	860
Right_Ventricle (E105)	Sigmoid_Colon (E106)	3573	2906
Sigmoid_Colon (E106)	Small_Intestine (E109)	1121	874
Small_Intestine (E109)	Thymus (E112)	3899	3163
Thymus (E112)	Spleen (E113)	2068	1628
Spleen (E113)	Penis_Foreskin_Keratinocyte (E058)	4588	3711

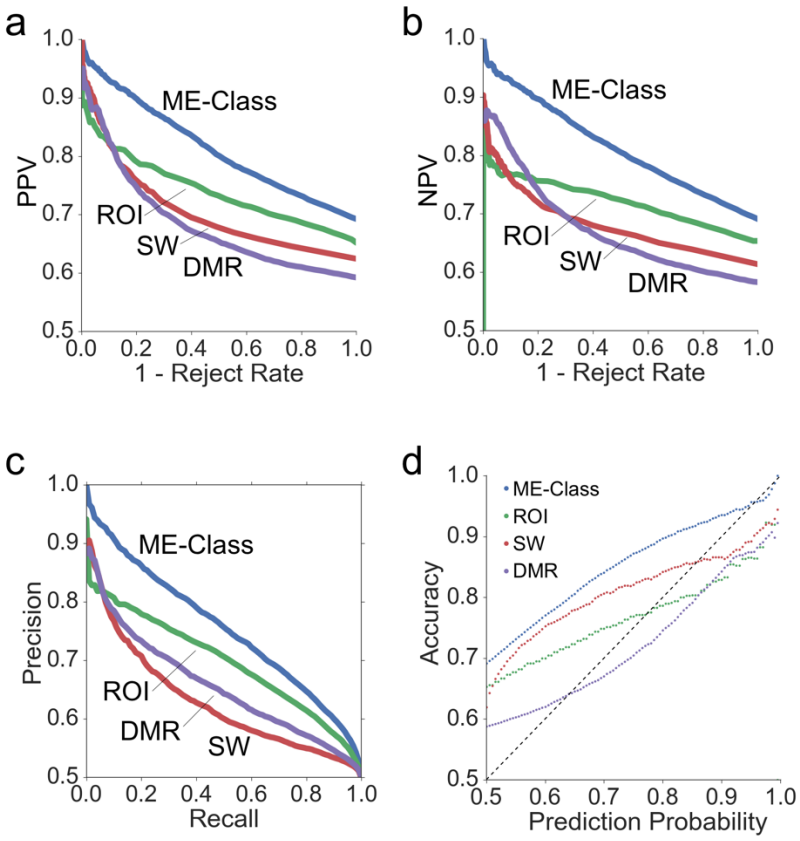


Supplementary Figure 3.1 Increasing the number of RF estimators from 100 to 1000 for the ROI classifier does not substantially increase performance as evaluated by: a) accuracy versus 1-reject rate, b) precision versus recall (PR AUC; 100 estimators: 0.70, 1000 estimators: 0.71), and c) ROC curve (ROC AUC; 100 estimators: 0.72, 1000 estimators: 0.73).

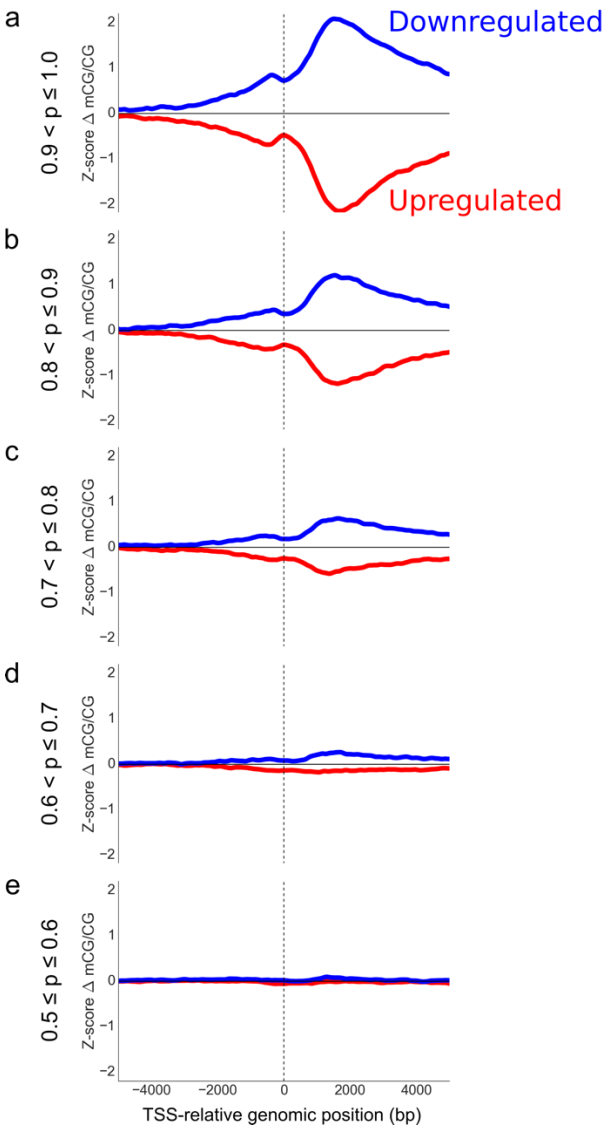


Supplementary Figure 3.2 Alternative full-gene methylation representations do not outperform TSS-centric representations. a) Heat map indicates methylation status at individual CpG sites – red is fully methylated, blue is fully unmethylated – for an example gene in two samples (Methyl. 1 and Methyl. 2). Individual points below indicate differential DNA methylation (Methyl. 2 – Methyl. 1) across the example gene at individual CpG sites. b) Whole Scaled Gene (WSG), c) Whole Gene (WG) and d) Uniform Gene Features (UGF) representation of the gene in (a). See additional description of each method in the Materials and Methods. Performance plots of TSS, WSG, WG, and UGF as reported by: e) accuracy versus 1-reject, f) ROC curve

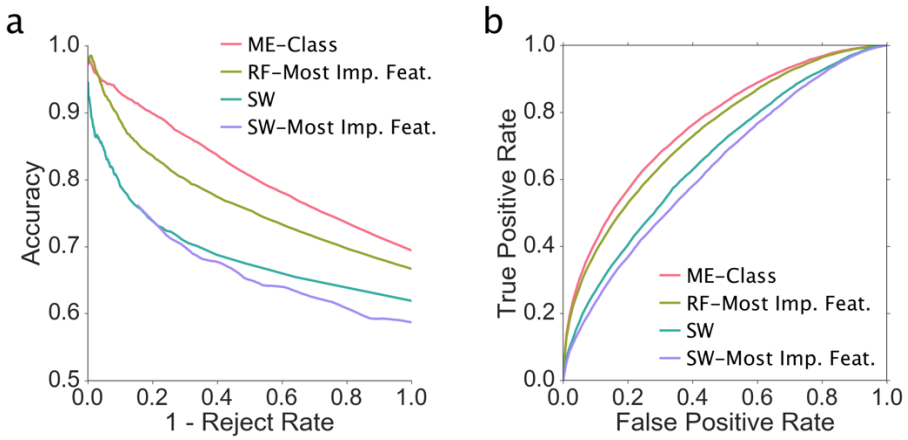
(ROC AUC; TSS: 0.76, WSG: 0.75, WG: 0.70, UGF: 0.65), and g) precision versus recall (PR AUC; TSS: 0.75, WSG: 0.74, WG: 0.69, UGF: 0.65). CGI = CpG island.



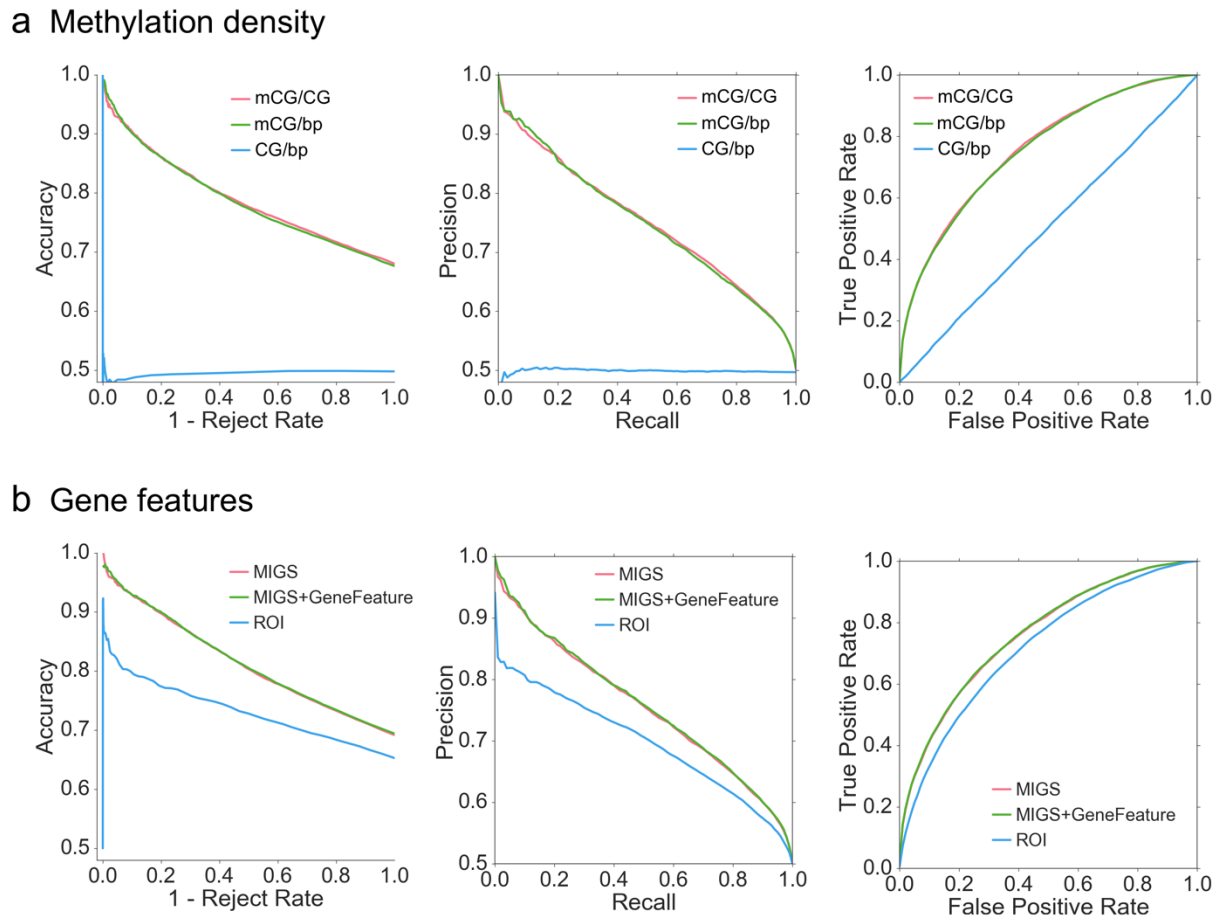
Supplementary Figure 3.3 Additional evaluation metrics for each method using 17 REP tissue differential samples: a) positive predictive value (PPV) versus 1- reject rate, b) negative predictive value (NPV) versus 1- reject rate, c) precision versus recall (PR AUC; ME-Class: 0.75, ROI: 0.70, DMR: 0.63, SW: 0.66) and d) accuracy versus the classifier's probability of prediction.



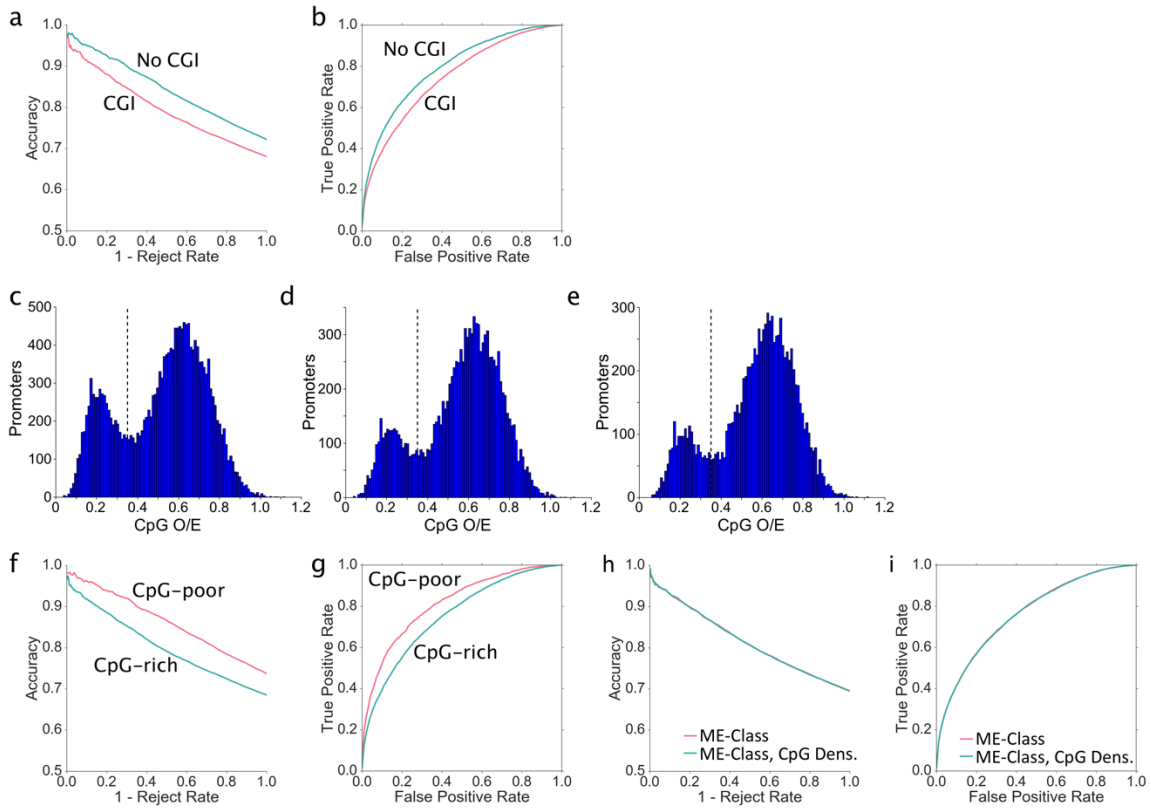
Supplementary Figure 3.4 Metagene plots of genes identified by ME-Class in REP data at different probabilities of prediction *p*. Blue curves represent the average Z-score normalized methylation difference between each sample for downregulated genes while red curves represent the average for upregulated genes.



Supplementary Figure 3.5 ME-Class outperforms classifiers using REP data based on only the most important methylation features, [+0.5kb, +2.5kb] around the TSS, as evaluated by: a) accuracy versus 1-reject rate, and b) ROC curve (ROC AUC; ME-Class: 0.76, RF-Most Imp. Feat.: 0.74, SW: 0.67, SW-Most Imp. Feat.: 0.64). RF-Most Imp. Feature is an ME-Class like classifier built using features from only the region [+0.5kb, +2.5kb] around the TSS. SW-Most Imp. Feat. is similar to the SW approach, but only using methylation from [+0.5kb, +2.5kb] around the TSS.

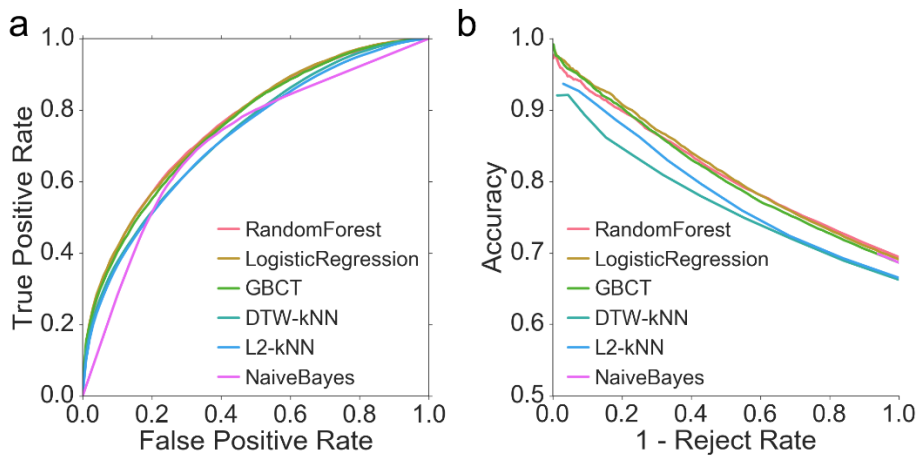


Supplementary Figure 3.6 The addition of methylated CpG density and gene body features (GF) does not increase ME-Class performance. a) Performance plots of ME-Class altered to use either mCG/CG, mCG/bp, or CpG density (200bp resolution, CG/bp) as input. (PR AUC; mCG/CG: 0.75, mCG/bp: 0.75, CG/bp: 0.50; ROC AUC; mCG/CG: 0.75, mCG/bp: 0.75, CG/bp: 0.50) b) Performance plots of ME-Class with and without adding gene body features (GF) from the ROI classifier including average internal exons, introns, and downstream features. ROI features are in Fig. 1d. (PR AUC; ME-Class: 0.75, ME-Class+GF: 0.76, ROI: 0.70; ROC AUC; ME-Class: 0.76, ME-Class+GF: 0.76, ROI: 0.72).

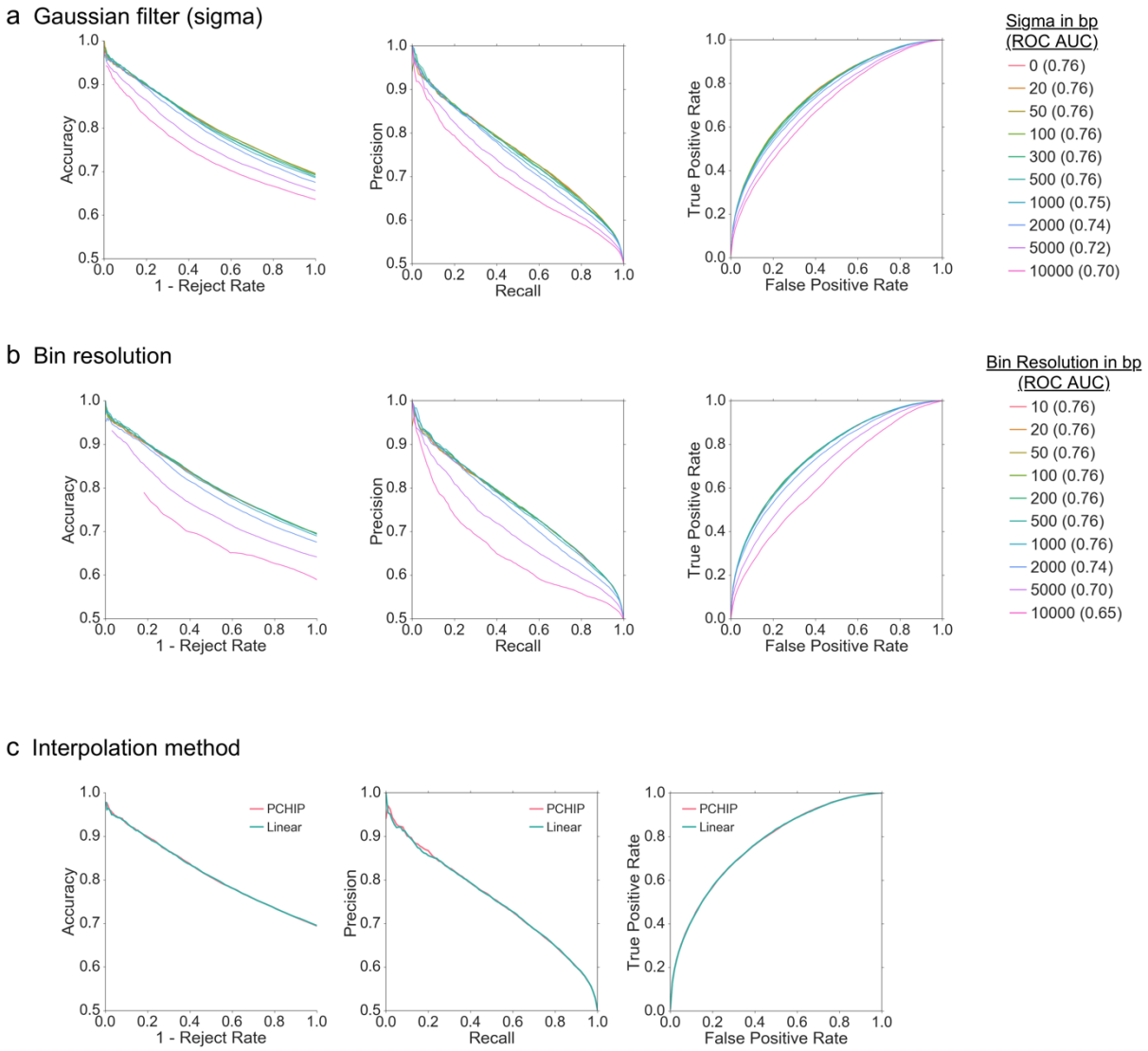


Supplementary Figure 3.7 CpG-poor genes are more predictive of expression classification. ME-Class performance for genes overlapping a CpG Island (CGI) by >=1bp is reported as: a) accuracy versus 1-rejection rate and b) ROC curve analysis (ROC AUC: No-CGI: 0.79; CGI: 0.75) c) Histogram of all genes with complete start and stop annotation according to RefSeq (n=19,175). Low CpG density genes comprise 26.0% (4,977 genes) while high CpG density genes comprise 74.0% (14,198 genes). d) Histogram of differentially expressed RefSeq genes (n=12,064 genes), where low CpG density genes comprise 18.8% (2,265 genes) while high CpG density genes comprise 81.2% (9,799 genes). e) Histogram of differentially expressed, interpolated RefSeq genes (n=10,524 genes) after applying our filtering parameters (see Materials and Methods). Low CpG density genes comprise 17.5% (1,842 genes) while high CpG density genes comprise 82.5% (8,681 genes). ME-Class performance is reported as: f) accuracy versus 1-rejection rate and g) ROC curve analysis (ROC AUC: CpG-poor: 0.8; CpG-rich: 0.75)

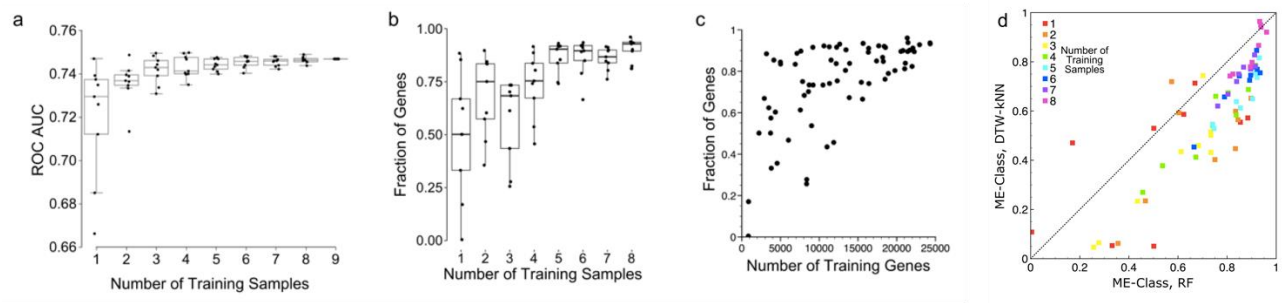
Cutoff between low and high CpG density genes at 0.35 observed/expected normalized CpGs +/- 1500bp of TSS. ME-Class performance with or without added feature of observed/expected normalized CpG density +/-1500bp of TSS is reported as h) accuracy versus 1-rejection rate and i) ROC curve analysis (ROC AUC: ME-Class: 0.76; ME-Class, CpG Density: 0.76).



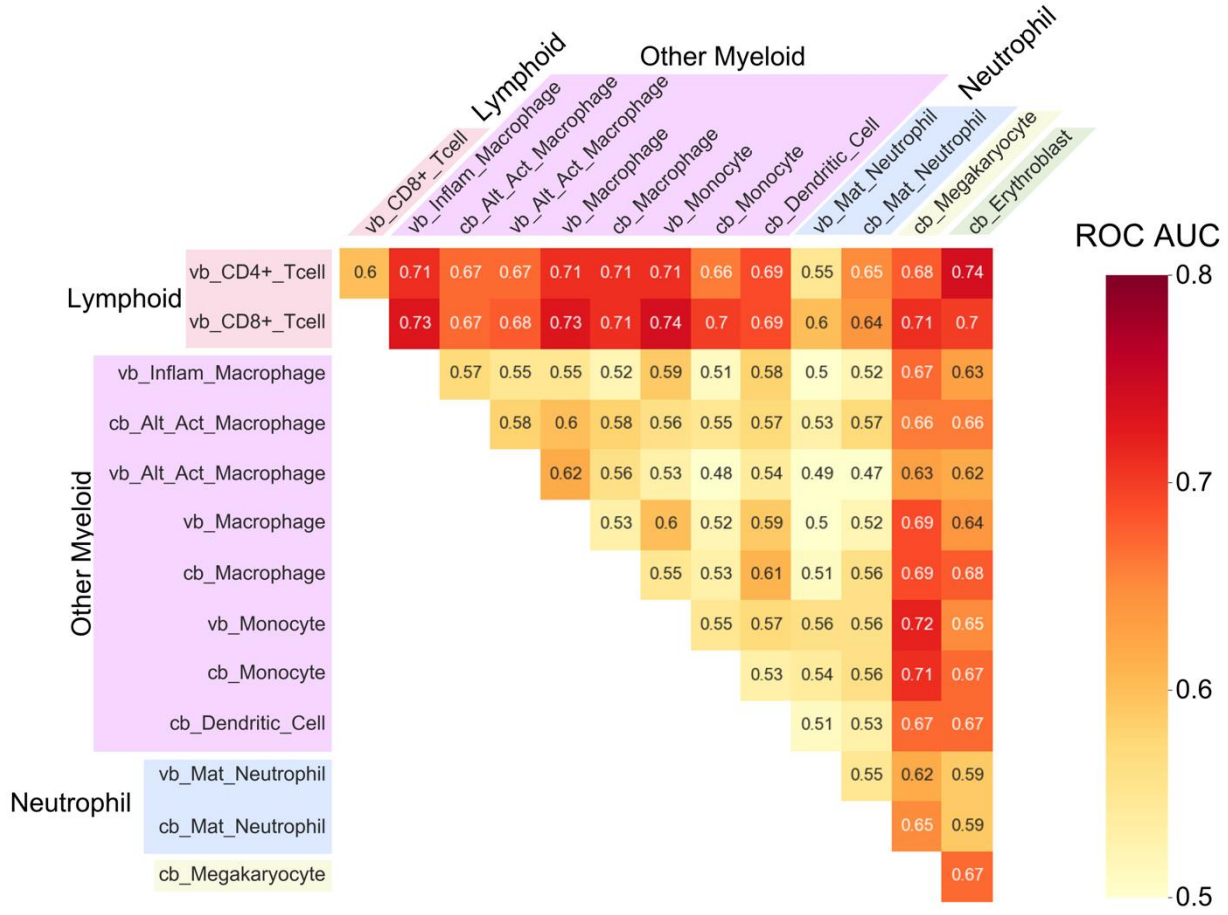
Supplementary Figure 3.8 Random Forest classifier performs similarly or outperforms alternatives based on classification performance as measured by a) ROC curve (ROC AUC; RF: 0.76, LR: 0.76, GBCT: 0.76, DTW-kNN: 0.73, L2-kNN: 0.73, Naïve Bayes: 0.71) and b) accuracy versus 1-reject rate. LR = Logistic Regression, GBCT= Gradient Boosted Classification Trees, DTW-kNN = Dynamic Time Warping based k-Nearest Neighbor, L2-kNN = Euclidean distance (L2) based k-Nearest Neighbor.



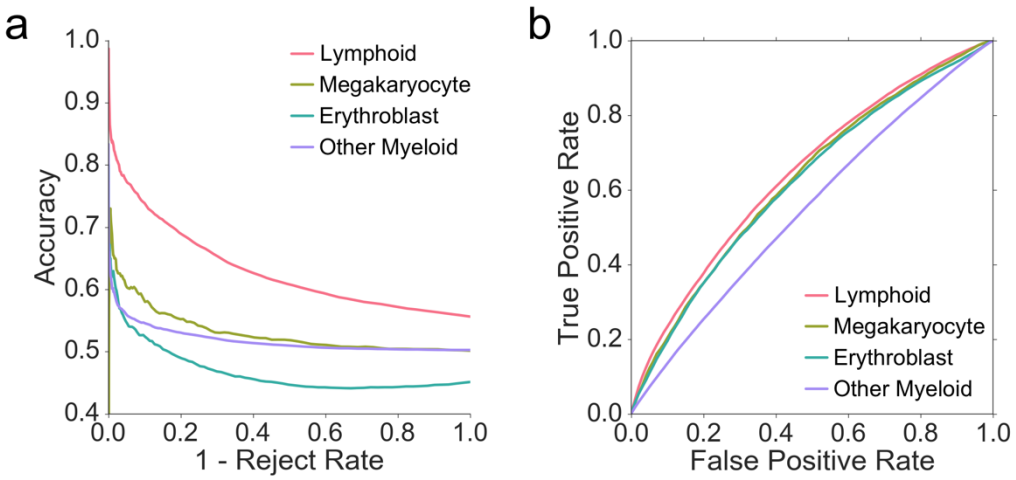
Supplementary Figure 3.9 Effect on ME-Class performance of tuning parameters for smoothing, bin resolution, and interpolation. Performance is reported as: accuracy versus 1-reject, precision versus recall, and ROC curve. a) Relationship between ME-Class performance and sigma for Gaussian smoothing with a constant bin resolution of 20bp. b) Relationship between ME-Class performance and the size of the bin resolution at a constant sigma of 50bp. c) Relationship between ME-Class performance and alternative interpolation method (PR AUC; PCHIP: 0.76, Linear: 0.76; ROC AUC; PCHIP: 0.76, Linear: 0.76).



Supplementary Figure 3.10 Number of training samples and genes determine ME-Class performance. The testing ROC AUC as a function of a) the number of training samples. The fraction of correctly identified genes using ME-Class with 9 evaluation samples as a function of b) the number of training samples and c) the total number of training genes from the training samples. In (a) and (b), each point indicates the performance across all genes in an individual sample comparison. In (c), each point indicates the number of training genes and fraction of genes returned for an individual sample comparison and set of training samples. Training genes in (c) is defined as the number of genes summed across all training samples. d) Fraction of correctly identified genes with 9 samples using ME-Class (DTW-kNN) versus ME-Class (RF). A gene can be counted multiple times if it shows up in multiple samples, although it will likely have different methylation profiles and expression values in each comparison. Permuted sets of all differential training samples (n=8) and a fixed set of differential evaluation samples (n=9) are randomly chosen from the REP dataset.



Supplementary Figure 3.11 Performance of Blueprint Epigenome samples using a similar leave-one-out differential sample evaluation cross-validation framework as used for the REP data (see Fig. 3.1e). The performance of ME-Class trained and evaluated solely using Blueprint samples is similar to that of a ME-class model trained from the REP dataset. Shown are ROC AUC of differential comparisons of randomly chosen single samples of each of the 14 cell types from Blueprint dataset.



Supplementary Figure 3.12 Performance of Blueprint neutrophil samples in comparison to other hematopoietic cell types. ME-Class is trained from the full REP dataset and performance is reported as: a) accuracy versus 1- reject rate and b) ROC curve analysis (ROC AUC; Lymphoid: 0.65, Megakayrocyte: 0.63, Erythoblast: 0.62, Other Myeloid: 0.55).

Chapter 4. Complex patterns of 5-methylcytosine and 5-hydroxymethylcytosine associate with gene expression changes in mammalian development and disease

This chapter consists of a manuscript in preparation with Manoj Singh and John R. Edwards.

4.1 Introduction

Endogenous human levels of 5-hydroxymethylcytosine (5hmC) have been detected in embryonic stem cells, neurons, liver, breast, testis, and placenta tissues at measurable global levels (~5-10% of CpGs) (166). Ten-eleven twelve (TET) enzymes have recently been found to be responsible for the oxidation of 5mC into 5hmC (22). TET enzymes further convert 5hmC to 5-formylcytosine (5fC) and 5-carboxylcytosine (5-caC). 5caC is thought to be removed though either thymine DNA glycosylase (TDG) in base excision repair (BER) or by decarboxylation (167). While 5mC is thought to passively be lost through incomplete copying of DNA methyltransferase 1 (DMNT1), TET enzymes imply a potential source of active demethylation. The deletion of TET enzymes result in viable murine offspring; however, progeny display varied developmental and epigenetic abnormalities. Murine *TET1/2* double KOs (DKOs) show global decreased levels of 5hmC and increased levels of 5mC (26). *TET1/2/3* triple KOs (TKOs) showed similar global levels of 5hmC depletion as well as poor differentiation into embryoid bodies and teratomas (27). *TET2* mutations and deletions have been identified as early genetic modifications in myelodysplastic syndrome (MDS) and acute myeloid leukemia (AML) (168), and these aberrations correspond with adverse clinical outcomes (169).

Single base pair resolution of 5hmC is unattainable through a standard 5mC assay of whole genome bisulfite sequencing (WGBS), since it unable to distinguish 5hmC from 5mC (170). Oxidative bisultifite sequencing (oxBS-seq) and *Tet* assisted bisulfite sequencing (TAB-seq) have been developed to assay 5hmC for single base-pair resolution quantification of 5hmC (28). An initial differential analysis of 5mC and 5hmC in murine brain development using TAB-seq showed that *TET2* activity is involved in marking active regulatory regions for demethylation

(39). It is unclear whether 5hmC is a transient mark in active demethylation, and global levels of 5hmC decrease quickly upon tissue culture which makes measurement of the modification reliant upon primary samples. While assayed samples of 5hmC are sparsely available in human tissues, Li et al. has recently compared matched normal and tumor samples of liver and lung (171). This study shows that 5hmC is observed as a marker of active transcription and associated with H3K4me1 at CpG island shores. However, experiments performed with genetic modifications of TET enzymes show conflicting results as to the effect on promoter 5hmC levels. Huang et al. show decreased 5hmC enrichment relative to normal mESCs for *Tet1* KDs, while *Tet2* KDs show a punctate increase of 5hmC 3' proximal to the TSS (40). Hon et al. shows decreased of 5hmC levels for both *Tet1/2* deletions at H3K4me3 promoters (41).

We hypothesize that tissue- and cancer-specific patterns of promoter 5hmC and 5mC can be categorized according to their contribution to expression change. We previously developed ME-Class to identify genes with a high probability of association between 5mC and gene expression (172). Here, we extend its functionality to incorporate 5hmC. We categorize highly predictive genes to elucidate similarly expressed gene signatures of promoter 5hmC and 5mC. This integrated analysis of promoter 5hmC and 5mC provides necessary information to understand the relationship of promoter 5hmC and 5mC to gene expression in mammalian development and disease.

4.2 Methods

4.2.1 TAB-seq, oxBS-seq, and RNA-seq

Mapped sequence reads for TAB-seq and RNA-seq in liver and lung tumor and matched normal samples were obtained from Li et al. (171). oxBS-seq and RNA-seq fetal and 6 week mouse brain samples were obtained from Lister et al. (39).

4.2.2 Estimation of 5mC and 5hmC levels

5mC and 5hmC levels were estimated using maximum likelihood methylation levels (MLML) (173) from either TAB-seq or oxBS-seq. MLML provides a simultaneous maximum likelihood based on binomial estimates of 5hmC and 5mC. Overshoot CpG sites refer to when combinations of oxBS-seq/TAB-seq and WGBS result in naïve estimates exceeding 0% or 100%. Conflict CpG sites correspond to sites where 5hmC and 5mC fall greatly outside their binomial confidence intervals. We used MLML with a significance level of α=0.05 for the binomial test at each CpG site and an expectation maximization convergence threshold of 1e-10. Counts of individual CpGs with estimated 5hmC and 5mC in all samples can be found in Supplementary Table 4.1.

4.2.3 Differential Expression from RNA-seq

We used featureCounts (174) to estimate feature counts over RefSeq reads. Differential comparisons were performed as defining differentially expressed genes as ($\log_2$(fold change) >=2). To create a standardized gene set with high quality methylation data, we excluded genes with ambiguous or incomplete TSS annotations, genes shorter than 5kb, genes with <40 CpGs assayed within +/-5kb of the TSS, genes where all CpGs within +/-5kb of the TSS had less than

0.2 methylation change, and alternative promoters. These filters were used to exclude non-coding and pseudogenes, genes shorter than the interpolation boundary, genes with low numbers of CpGs to reduce bias caused by individual CpGs, and genes with no methylation changes at their promoter, respectively. We only included RefSeq genes with cdsStartStat and cdsEndStat with 'cmpl' according to the UCSC Table Browser. For any RefSeq genes with multiple RefSeq IDs corresponding to the same TSS location, we used a single RefSeq ID with the lowest accession number and excluded the remainder. This is a conservative method to simplify the annotations of genes with alternative promoter annotations. A full summary of differentially expressed filtered gene counts can be found in Supplementary Table 4.2.

4.2.4 5hmC incorporation in ME-Class

MLML produces an estimate of 5hmC and 5hmC for each CpG site. ME-Class was extended to independently interpolate 5mC and 5hmC data using PCHIP interpolation and Gaussian smoothing (50bp bandwidth) as described in Schlosberg et al. (172). Δ5hmC/CG and Δ5mC/CG interpolated curves were subset to 20bp to create feature vectors for classification. Δ5mC/CG corresponds to the 5mC feature vector and Δ5hmC/CG corresponds to the 5hmC feature vector. Δ5mC/CG+Δ5hmC/CG corresponds to using both the 5mC and 5hmC feature vectors for the classification.

4.2.6 Evaluation Framework

All evaluation frameworks are established as described in Schlosberg et al. (172) with the following exceptions. The fetal-6 week comparison of mouse brain development used an intra-sample 10-fold cross validation as this is currently the only sample in mouse with single base

pair resolution of 5hmC and 5mC. The normal liver-lung comparisons kept examples of the same tissue type, but still excluded the testing genes from each fold of the evaluation set because this dataset represents the only single base pair resolution dataset in human.

4.2.5 Unsupervised Clustering of 5hmC and 5mC

Unsupervised hierarchical agglomerative clustering (complete linkage) was performed on Δ5mC/CG and Δ5hmC/CG in the region [+0.5kb,+2.5kb] for the TSS for correctly predicted genes from ME-Class. Subsetting our predictions required setting a working threshold for the probability of prediction. The sample size of differentially expressed genes in this study is smaller than used in Schlosberg et al. (172). Therefore, we set the following range of probabilities of prediction for each experiment: [0.7,1.0] fetal-6wk mouse, [0.8,1.0] normal liver-tumor, [0.7,1.0] normal-tumor liver and lung. In the metagene plots of unsupervised results Δ5mC/CG+Δ5hmC/CG corresponds to the summation of curves.

4.3 Results

4.3.1 ME-Class identifies 5hmC and 5mC signatures in mouse brain development

ME-Class was extended to interrogate the added information that 5hmC provides to predict differential gene expression change between fetal and 6-week mouse brain development. We demonstrate good performance to predict gene expression change from 10-fold cross validation (Figure 4.1a: ROC AUC: Δ5mC/CG; Δ5hmC/CG: 0.73, Δ5mC/CG: 0.70, Δ5hmC/CG: 0.67). In mammalian brain development, 5hmC increases the accuracy of ME-Class as a function of the rejection rate (Figure 4.1b). We sought to understand what patterns of 5hmC and 5mC contribute to this increase in ME-Class performance.

We conduct a post-hoc unsupervised clustering analysis of identified signatures of 5hmC and 5mC that associate with expression change. We observe three distinct classes of differential 5hmC and 5mC signatures and provide gene examples of each pattern in Supplementary Figure 4.2. In Figure 4.1c, we observe a net decrease in methylation 3' proximal to the TSS and a corresponding increase in expression, which is an observable signature in ME-Class with only 5mC information. The converse is shown in Figure 4.1d, where there is a net increase in 3' proximal TSS methylation and a decrease in expression. We obtain added knowledge in Figure 4.1e, where we show a net neutral change in Δ5mC/CG+Δ5hmC/CG 3' proximal to the TSS with a decrease in expression. This cluster of 5hmC and 5mC gene signature implies that during conversion from 5mC to 5hmC the gene is preferentially downregulated. Increased ME-Class performance from using Δ5mC/CG+Δ5hmC/CG versus Δ5mC/CG alone is most likely due to subtle combinations of 5hmC and 5mC.

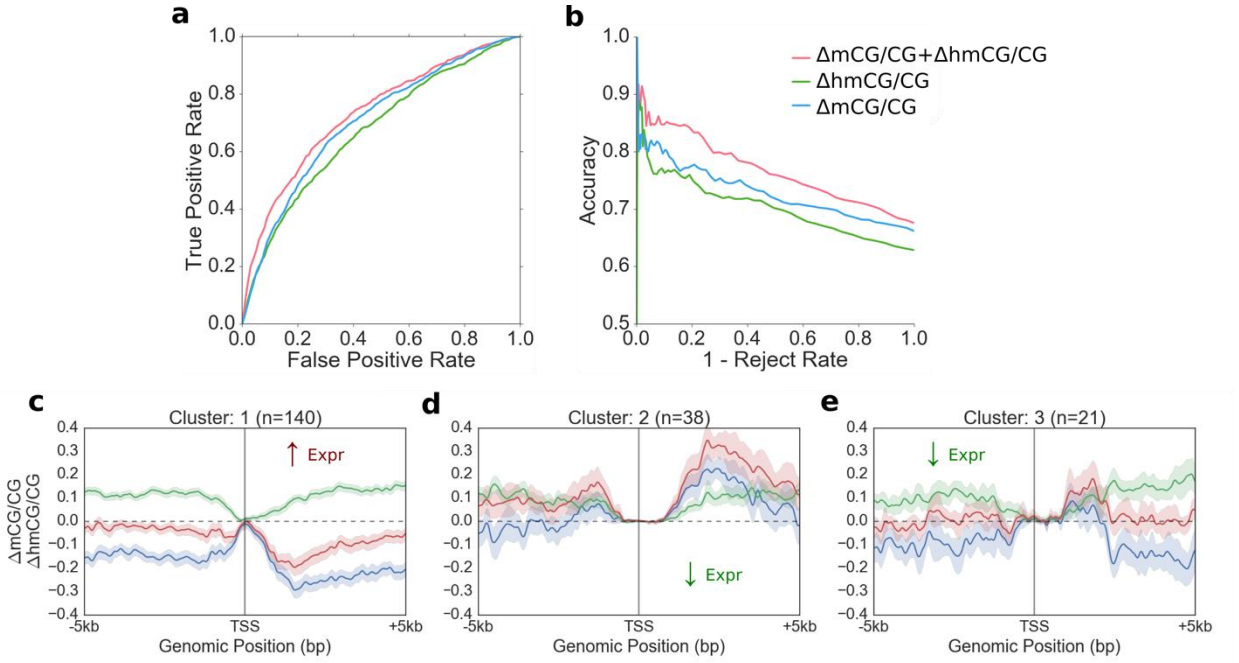


Figure 4.1 ME-Class identifies 5hmC and 5mC signatures in mouse brain development. a) ROC AUC analysis: Δ5mC/CG; Δ5hmC/CG: 0.73, Δ5mC/CG: 0.70, Δ5hmC/CG: 0.67 b) accuracy versus 1-reject rate. Metagene plots for cluster of genes with: c) increased expression with corresponding decreased TSS 3' proximal 5mC (n=140), d) decreased expression with increased TSS 3' proximal 5mC (n=38), and e) decreased expression with coordinated TSS 3'proximal decrease in 5mC and increase in 5hmC (n=21). Shading in metagene plots indicates the 68% bootstrapped confidence interval.

4.3.2 Human tissue-specific patterns of 5hmC and 5mC

We next sought to determine whether murine, tissue-specific model perform similarly to those trained in human samples of liver and lung. We examined differential comparisons between human liver and lung and evaluated using leave-one-sample-out framework, and show that both upregulated and downregulated genes are similarly represented (Supplementary Table 4.2). ROC AUC analysis indicated that 5mC and 5hmC information only marginally increases the ability to predict differential expression class (Figure 4.2a; ROC AUC: 5mC;5hmC: 0.84, 5mC: 0.83, 5hmC: 0.7). Differential human tissue-specific comparisons also show a large difference in performance of the predictive ability between 5mC and 5hmC. We also demonstrate equivalent performance using both 5hmC and 5mC as using only 5mC alone by examining the accuracy versus the rejection rate (Figure 4.2b).

To understand why 5hmC does not contribute to increasing ME-Class performance in these human tissues, we performed an unsupervised analysis of highly predicted genes. In Figure 4.2c, we observe a large cluster of hypermethylated genes with decreased expression (n=1,333). Genes

with increased expression were more difficult to predict, however, we identified a sizeable cluster of TSS 3' proximal hypomethylated genes with increased expression (Figure 4.2d, n=276). We did not observe a cluster of 5mC to 5hmC conversion as we did in the model of mouse development. The pattern of differential 5hmC and 5mC closely follows that of 5mC alone. Also, the overall of performance of ME-Class with 5hmC and 5mC is not increased using 5hmC. This implies that since 5hmC exists at relatively low levels in liver (2.27%-5.68%) and lung (1.94-3.04%) (171), 5hmC might not play as important a role as it does in brain development (17.2%) (39).

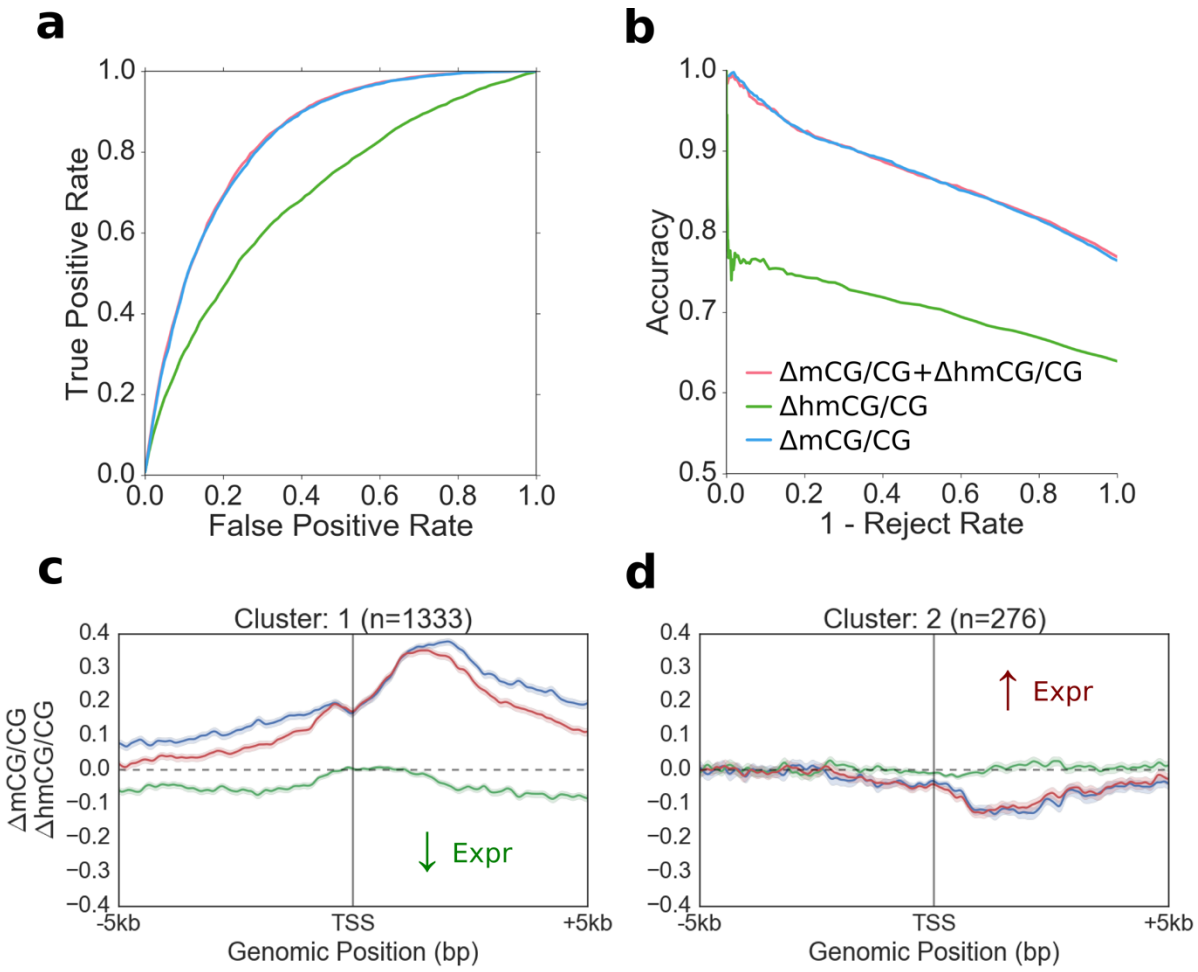


Figure 4.2 ME-Class performance in human tissue-specific model (normal liver-lung). a) ROC AUC analysis: 5mC/CG; 5hmC/CG: 0.84, 5mC/CG 0.83, 5hmC/CG: 0.7 b) accuracy versus 1-reject rate. Metagene plots for c) cluster of hypermethylated genes with decreased expression (n=1,333) and d) cluster of TSS 3' proximal hypomethylated genes with increased expression (n=276). Shading in metagene plots indicates the 68% bootstrapped confidence interval.

4.3.3 Diversity of human cancer-specific patterns of 5hmC and 5mC

We next sought out to compare how these patterns of 5hmC change in tumorigenesis. We compared differential comparisons of liver and lung cancer versus their normal controls in a leave-one-sample-out framework. We observe that differentially expressed genes are preferentially downregulated in normal versus tumor samples (Supplementary Table 4.2). We show moderate performance averaged over the five differential samples examined (Figure 4.3a: ROC AUC: Δ5mC/CG; Δ5hmC/CG: 0.72, Δ5mC/CG: 0.72, Δ5hmC/CG: 0.59). There exists a large difference in predictive ability between 5mC and 5hmC, implying that 5hmC is insufficient to alone predict expression change in cancer-specific samples. A similar behavior was also observed when examining the curve of accuracy versus 1-reject rate (Figure 4.3b). However, we observe that two samples (lung normal-tumors samples 1,2) perform significantly worse than the remaining samples (Supplementary Figure 4.2).

Unsupervised clustering of highly predicted cancer-specific genes uncovers a diversity of 5hmC and 5mC patterns. We observe a cluster of TSS 3' proximally demethylated genes set within a larger hypomethylation region with an increase in expression (Figure 4.3c, n=414 genes). There exists a cluster of TSS 3' proximal methylated genes with a decrease in expression (Figure 4.3d,

n=229 genes). We also notice a cluster of downregulated, TSS methylated genes (Figure 4.3e, n=319). In each of these patterns, the summation of 5hmC and 5mC closely follows the 5mC pattern, and 5hmC is primarily constant throughout. These patterns resemble those found in cancer cell lines (96). This implies that promoter 5mC rather than 5hmC is primarily associated with gene expression change in cancer, which is consistent with a decrease in 5hmC from normal liver (2.27%-5.68%) and lung (1.94-3.04%) to tumorigenic liver (0.7-2.07%) and lung (0.65-1.07%), respectively (171).

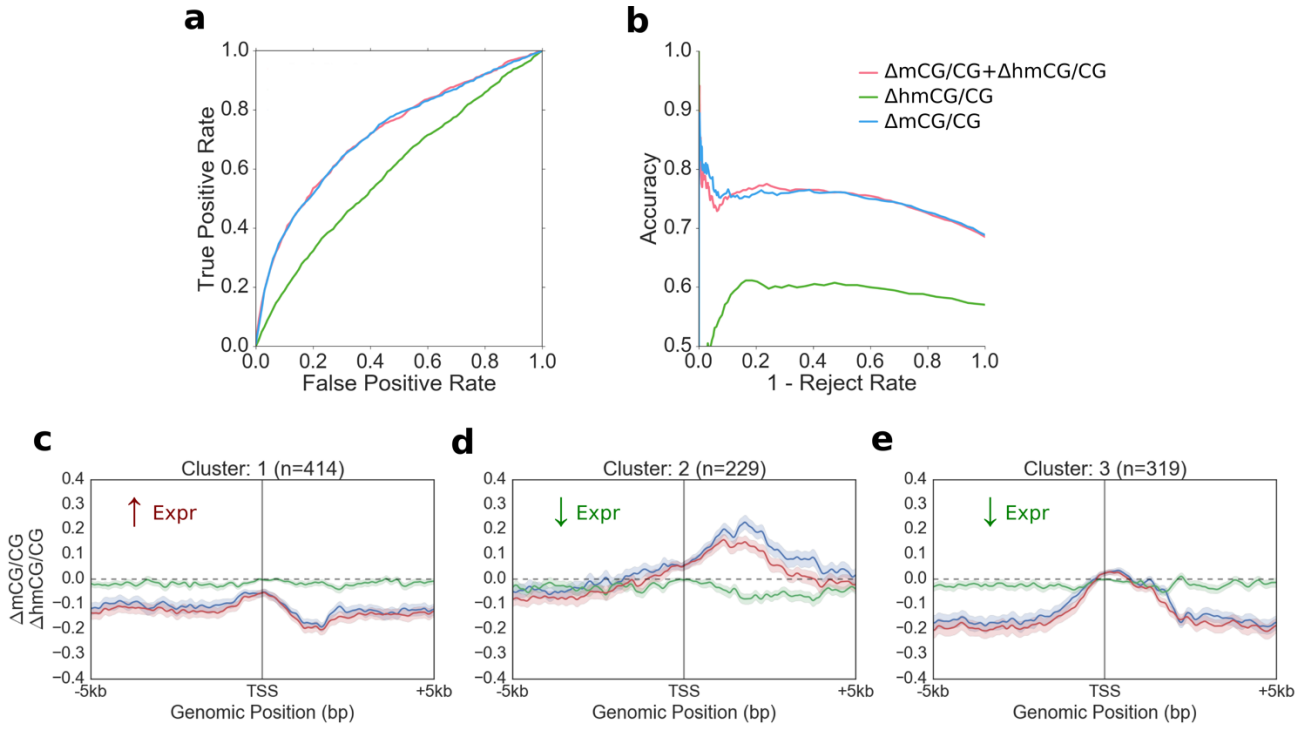


Figure 4.3 ME-Class performance and pattern diversity of 5hmC and 5mC in human cancer-specific model. a) ROC AUC analysis: 5mC/CG; 5hmC/CG: 0.72, 5mC/CG: 0.72, 5hmC/CG: 0.59 b) accuracy versus 1-reject rate. Metagene plots for: c) cluster of TSS 3' proximally demethylated genes with an increase in expression (n=414 genes) d) cluster of TSS 3' proximal methylated genes with a decrease in expression (n=229 genes) e) cluster of TSS methylated

genes with a decrease in expression (n=319). Shading in metagene plots indicates the 68% bootstrapped confidence interval.

4.4 Discussion

We successfully extended ME-Class to predict gene expression classification from both 5hmC and 5mC. We observe similar levels of performance in both tissue- and cancer-specific differential comparisons for the combination of 5hmC and 5mC as we do to using 5mC alone. We also identify a class of conversion between 5mC and 5hmC in a model of mouse brain development corresponding to a downregulation in gene expression. This observation in a highly predictive subset of the differentially expressed genes is contrary to observations by examining all genes as reported in Lister et al. (39). In this study, 5hmC is shown to be enriched in the gene body and primarily associated with genes of high expression. We identify that conversion of 5mC to 5hmC primarily is associated with the downregulation of gene expression, which we speculate could be due to the recruitment of additional silencing factors. We also observe that promoter 5hmC and 5mC patterns are unique in brain development but not in difference between the liver and lung or cancer and normal. 5hmC could be a possible source of gene regulation by increasing the plasticity of gene expression in brain tissue. However, these observations are confounded by potential sources of clonal heterogeneity since these are bulk tissue samples. We would propose to confirm this novel class of 5mC conversion through cell sorted populations of neurons, astrocytes, and oligodendrocytes.

In the quantification of 5hmC and 5mC, MLML estimated lower overshoot and conflict CpGs from TAB-seq than from oxBS-seq. This most likely occurs because oxBS-seq requires two

bisulfite sequencing experiments while TAB-seq only requires one, therefore, increasing the probability of error (173). One limitation of current work is the lack of high resolution 5hmC datasets poses a significant problem for the establishment of proper training and testing paradigms to evaluate genes with associations between 5hmC, 5mC, and gene expression. We made conservative assumptions in this study to exclude identical genes, but we could not exclude samples of the same tissue type in all comparisons. Prior work from Chapter 3 indicates that we need three differential samples to improve performance. However, we do not feel these models are overfit since we are not comparing to identical tissue samples.

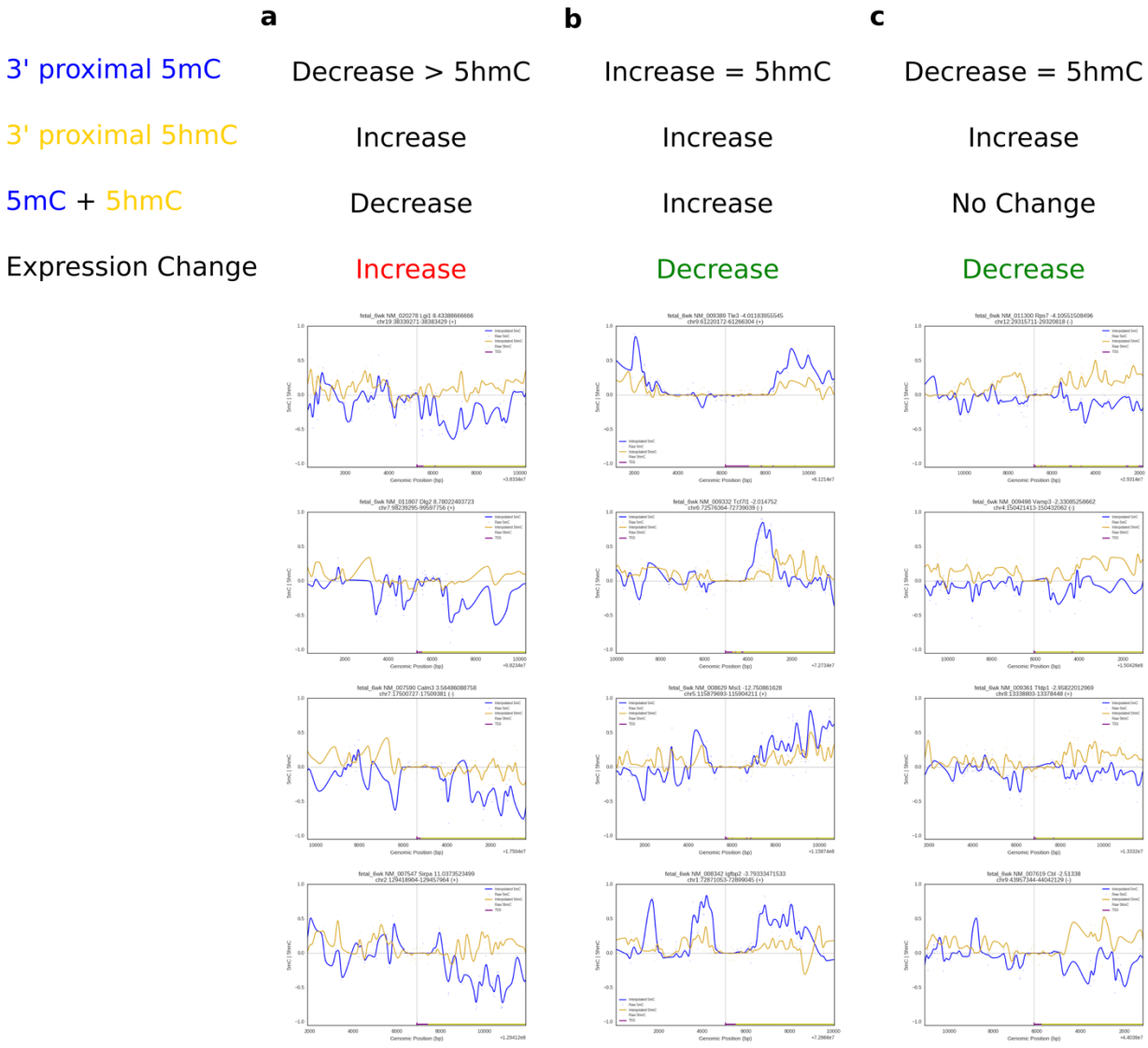
Another difficulty in the analysis of 5hmC is the relative timescale at which 5hmC is assayed. 5hmC is an intermediate cytosine modification which is not replicated over mitosis. Lack of gene expression correlating 5hmC patterns in normal lung and liver may be because these are non-senescent tissue samples as compared to neuronal tissue in mouse development. The relative scarcity of 5hmC in the genome relative to 5mC might also explain its lack of predictive ability for expression class change. We encourage increasing the number of high resolution 5hmC and 5mC datasets to study the relationship between 5hmC, 5mC, and gene expression in each tissue where 5hmC is measurable and especially in brain tissue subtypes.

Supplementary Table 4.1 Single sample CpG counts with MLML estimated CpGs. T=tumor, N=matched normal.

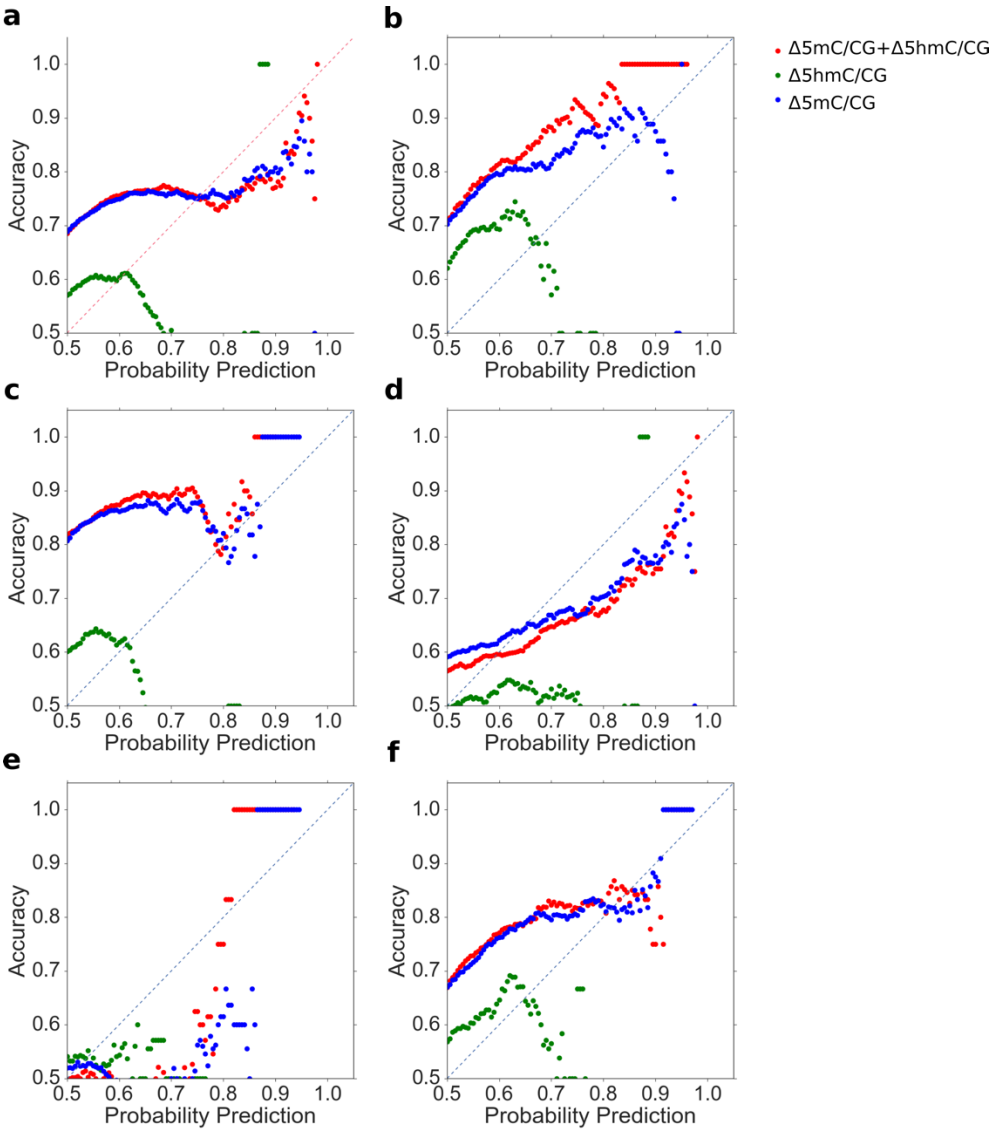
	Design			*MLML Statistics*			
Study	**5hmC Assay**	**Organism**	**Sample**	**Overshoot**	**Conflicts**	**Accepted CpGs**	**CpG Coverage**
Li et al. *Genome Res* 2016	oxBS-seq	Human	Liver_N1	7483717	45698	23373853	0.828347517
	oxBS-seq	Human	Liver_N2	7779952	35689	23281047	0.82505856
	oxBS-seq	Human	Liver_N3	8830879	63528	23395102	0.829100562
	oxBS-seq	Human	Liver_T1	8706959	56985	23027265	0.816064762
	oxBS-seq	Human	Liver_T2	9048391	63705	24285424	0.860652742
	oxBS-seq	Human	Liver_T3	9388948	71465	23380024	0.828566212
	oxBS-seq	Human	Lung_N1	9487853	64312	24490004	0.867902866
	oxBS-seq	Human	Lung_N2	8515568	53325	23425058	0.830162175
	oxBS-seq	Human	Lung_N3	8184630	28879	23159840	0.820763097
	oxBS-seq	Human	Lung_T1	10236654	94104	24675947	0.874492513
	oxBS-seq	Human	Lung_T2	9656229	84745	23617849	0.836994508
	oxBS-seq	Human	Lung_T3	9529502	47862	23186713	0.821715451
Lister et al. *Science* 2013	TAB-seq	Mouse	Fetal_Brain	250772	1	19634599	0.919964499
	TAB-seq	Mouse	6wk_Brain	242633	5	19875929	0.931271837

Supplementary Table 4.2 Differentially expressed gene counts for 5hmC datasets

Organism	Sample 1	Sample 2	Genes (FC<=-2,FC>=2)	Genes (FC<=-2)	Genes (FC>=2)	Genes Down (%)	Genes Up (%)
Human	Liver_N1	Lung_N1	3456	1554	1902	44.97%	55.03%
Human	Liver_N1	Lung_N2	3677	1567	2110	42.62%	57.38%
Human	Liver_N1	Lung_N3	3562	1622	1940	45.54%	54.46%
Human	Liver_N2	Lung_N1	2771	1242	1529	44.82%	55.18%
Human	Liver_N2	Lung_N2	3275	1494	1781	45.62%	54.38%
Human	Liver_N2	Lung_N3	3077	1541	1536	50.08%	49.92%
Human	Liver_N3	Lung_N1	2179	1150	1029	52.78%	47.22%
Human	Liver_N3	Lung_N2	2542	1224	1318	48.15%	51.85%
Human	Liver_N3	Lung_N3	2544	1339	1205	52.63%	47.37%
Human	Liver_N1	Liver_T1	1855	1061	794	57.20%	42.80%
Human	Liver_N2	Liver_T2	2550	1926	624	75.53%	24.47%
Human	Lung_N1	Lung_T1	2196	1270	926	57.83%	42.17%
Human	Lung_N2	Lung_T2	403	250	153	62.03%	37.97%
Human	Lung_N3	Lung_T3	2767	1559	1208	56.34%	43.66%
Mouse	Fetal_Brain	6wk_Brain	3654	2063	1591	56.46%	43.54%



Supplementary Figure 4.1 ME-Class results identifies unique classes of 5hmC and 5mC in mammalian brain development. Differential 5hmC and 5mC signatures of genes with: a) increased expression with corresponding decreased 5mC (see Figure 4.1c for metagene), b) decreased expression with increased 5mC (see Figure 4.1d for metagene), and c) decreased expression with coordinated decrease in 5mC and increase in 5hmC (see Figure 4.1e for metagene8).



Supplemental Figure 4.2 Tumor-normal human liver and lung differential comparison performance. Plotted is accuracy of gene results versus the prediction of probability for a) average of all five tumor-normal samples, b) liver_T1-liver_N1, c) liver_T2-liver_N2, d) lung_T1-lung_N1, e) lung_T2-lung_N2 f) lung_T3-lung_N3. T=tumor, N=matched normal.

Chapter 5. Conclusions and Future Directions

Cell type specific expression profiles are defined by sequence modifications that affect transcription factor binding. DNA methylation is a heritable modification that increases the content of sequence information. However, its role as a regulator of gene expression has primarily been attributed to X-inactivation, silencing of transposable elements, and imprinting. The goal of this dissertation is to identify the contribution that DNA methylation contributes to gene expression in the context of development and disease. From genome-wide methylation data, we provide a straightforward method to identify amplifications and deletions and biologically confirm the existence of a large amplification in both *in vitro* and *in vivo* models. We then developed a comprehensive, alternative approach (ME-Class) to provide a standard integrated analysis of DNA methylation and gene expression changes and to identify putatively functional genes. These studies provide stronger evidence for differential methylation to define tissue- and cancer-specific transcriptional changes.

5.1 DNA Methylation and Copy Number Variation

We used DNA methylation data to screen multiple breast cancer cell lines to determine if CNVs could serve as biomarkers for response to a model of long term estrogen deprivation therapy. We identified *ESR1* as the most amplified region in a model of LTED in breast cancer and it is associated with poor clinical outcomes. This implies that ER receptor is amplified to scavenge for estrogen in an estrogen deprived environment, as observed in common breast cancer therapies. We also observe a decrease in this amplification upon the recovered state. There also exists a paradoxical effect of estradiol as a late-stage endocrine therapy in advanced ER+ breast cancers. A primary difficulty in the CNV analysis was the lack of high resolution genomic data, such as whole genome sequencing, from which to estimate a background distribution for

normalization (124). The Methyl-MAPS analysis pipeline (36), which relies on read frequencies from both methylated and unmethylated sequencing compartments, could be improved to include Bayesian probabilities for single base pair methylation estimates to account for biases in each compartment before normalization. With better estimation of CNVs, DNA methylation could also be used as a marker of genetic drift for cell line analysis in comparison to PDX models for tracking changes relative to reference tumor genomes (123). However, the plummeting cost of sequencing and the high false negative rate of our approach makes it advantageous to pursue established methods of CNV identification through whole genome sequencing and genome segmentation algorithms.

5.2 DNA Methylation and Differential Expression

We provide a systematic model, training, and evaluation scheme to assess the ability of using differential WGBS methylation data to predict differential expression. The ability to understand the contribution that differential methylation plays in expression change is vital to predicting the functional effects of induced methylation changes. Induced methylation and demethylation experiments through DMNT- or TET-based zinc fingers proteins or CRISPR technologies have proved invaluable to demonstrating that methylation patterns are functional (59-61). However, these experiments have not assumed any specific correlation between methylation and gene expression. I hypothesize that disease states with heterogeneous endogenous chromatin environments will respond differently to induced methylation and demethylation. I propose that DNA methylation acts differently in each context. In normal tissues, inactive lamin associated domains (LADs) and large organized chromatin lysine modifications (LOCKs) exist in heterochromatin. DNA methylation changes might not have much effect on expression changes

since LADs and LOCKs exist in closed chromatin environments. In active transcriptional factories and gene-dense topologically associated domains (TADs), chromatin is accessible and DNA methylation is likely to affect gene expression change. Bivalent, poised chromatin regions (marked by H3K4me3 and H3K27me3) also exist with accessible chromatin regions, and induced DNA methylation changes would also be likely to affect gene expression. ME-Class predictions would be particularly interesting to test induced methylation technologies in cancer where LOCKs are commonly disrupted and functional DNA methylation-induced gene silencing has been observed (175). In the analysis of REP datasets, ME-Class also identifies CpG-poor (or no CpG Island) genes as more predictive than CpG-rich (with CpG Island) genes. CpG-poor genes might be a more likely target for induced methylation changes. For example, induced, targeted DNA methylation of oncogenes with no associated CpG-island such as *KRAS* or *TLR4* might have a greater effect on expression than a CpG-island associated oncogene like *MYC* or *VEGF*. If we can identify which genes are like to have their expression changed by DNA methylation, this opens the door to specifically targeted methylation therapies as an alternative to the current clinically approved global demethylation agents.

The difference in the average differential methylation signatures between expression classes in the [+0.5kb,+2.5kb] region of the TSS shows a clear correlation with the performance of ME-Class. Each of the samples that we used to test ME-Class is obtained from a bulk average of individual cells. In addition, complex tissues such as the brain consists of multiple classes of cells (neurons, astrocytes, and oligodendrocytes), each of which are thought to have cell type specific DNA methylation patterns (176). In a single cell, DNA methylation exists as a binary signal and CpG hemimethylation is rare. Therefore, we cannot exclude cell population

heterogeneity as a source of this correlation in the 3' proximal region. Single cell sequencing will be a useful tool in understanding the significance of these putative methylation patterns (177), however, these methods are currently limited by reduced library complexity (178). However, there are emerging technologies to create bulk sample WGBS libraries from lower input quantities (179).

Differential TSS 3' proximal methylation is consistent with a previously proposed model where p300 binds at downstream unmethylated CpG Islands (85), thus increasing expression of the corresponding genes. Previous studies have also observed decreased downstream methylation correlating with increases in the H3K4me3 active mark that skew downstream of the TSS (163, 164). Both observations are consistent with the hypothesis that this downstream region is important for prediction of differential gene expression.

DNA methylation is proposed to have a role not only in transcriptional regulation but also in alternative splicing, regulated through at a subset of exons either through CTCF and MeCP2, which adjusts the elongation rate of PolII, or HP1, which recruits splicing factors (180). It is possible to use ME-Class (without modification) to examine alternative promoters and isoforms using isoform level RNA-seq data. However, there are technical challenges to this analysis. Isoform level analyses still display high levels of noise, especially for low levels of expression (181). This could be alleviated by predicting isoform expression rank rather than absolute expression levels. Another difficulty would be to manage observing overlapping copies of isoforms that originate from the same DNA methylation signature for a given gene. This could

be mitigated by binning and scaling transcripts according to length, and adhering to the training, testing, evaluation framework that we have established in ME-Class.

5.3 DNA Methylation and Enhancers

Transcription factors coordinately bind to enhancers to modulate gene expression. DNA methylation can alter transcription factor binding affinities, thus impacting gene expression (182). Enhancer regions with variable methylation have been identified to correlate methylation with gene expression (30). However, the relative importance of promoter versus enhancer methylation information is unclear. A previous study concluded that enhancer methylation is most important for association with gene expression in cancer from ChromHMM annotations of enhancer sites (183). I hypothesize that a comprehensive model of DNA methylation at a given gene's promoter and enhancer regions would lead to a predictive understanding of how DNA methylation controls gene expression. A requirement for this analysis would be to ascribe specific enhancers to promoters, however, this task is an open area of research (184). Enhancers primarily act in a cis-regulatory mechanism within 100kb of the gene's TSS. Therefore, DNA methylation surrounding all experimentally verified enhancer regions, as verified by the Vista Enhancer Browser (185), within 100kb of a given gene's TSS can be transformed into a feature vector of average enhancer differential methylation. This feature vector can then be used as input for classification or regression of differential gene expression. This analysis would be useful to examine the relative impact of TSS proximal methylation changes versus those at enhancers using feature importance.

5.4 DNA Methylation and Histone Modifications

DNA methylation is a stable covalent modification which adds further information to sequence to provide information about gene expression. In FFPE samples, histone modifications and RNA are rapidly degraded by proteases and ribonucleases, and therefore cannot be used transcriptional profiles of disease states. In hematopoiesis, ME-Class identified that DNA methylation-based changes are more predictive of expression change in differences between the early myeloid and lymphoid lineages. This is consistent with a hypothesis that these predictive DNA methylation changes occur early in blood development (20). DNA methylation changes have also been observed to regulate CpG dense genes early in development through constitutive promoter hypomethylation (186). These observations imply that in normal development, DNA methylation acts as a lock on predetermined transcriptional events. Histone modifications have been used to predict gene expression in single samples (187, 188). In normal development, histone modifications provide genes more fine-tuned control over gene expression than DNA methylation (42). The presence of substantially more histone modifications than DNA methylation provides for a theoretical increase in the combinatorial information that exists in histone modifications.

5.5 DNA Methylation Representation

We observe a shift in the feature importance of our classifier from [+0.5kb,+2.5kb] in differences between tissues (REP) to [-0.5kb,+1.5kb] in cancer (TCGA). This suggests that DNA methylation signatures are fundamentally different with respect to tissue- versus cancer-specific expression changes. To confirm this hypothesis, I would assay tumor-normal paired WGBS and RNA-seq datasets and perform post hoc unsupervised clustering on ME-Class predicted genes in

comparison to REP genes. I would anticipate an increase in the ratio of genes with TSS-centric patterns in cancer relative to 3' proximal patterns in the REP dataset.

As machine learning grows in popularity in biology, proper establishment of training and testing paradigms will be vital given that biological data is often limited in sample size due to the high price of obtaining quality data. Even observing patterns of differential DNA methylation can be prone to memorization and overfitting. A prior examined testing framework used a leave-one-differential-sample-out cross validation by comparing H1 hESC to each of the normal tissues in the Roadmap Epigenomics Project. Unsurprisingly, we observed nearly perfect classification accuracy since methylation patterns across differentiated tissues are remarkably similar. This experiment reinforced the importance of excluding any data from any observed samples in the training set. We also commonly observe that the nearest neighbor of the testing differential methylation curve was often the same gene in a different training sample, which stresses the importance of excluding any examples of the same gene from the training dataset.

ME-Class also observed that lower resolution bins (up to 200bp) can accurately capture the complexity of the methylation signal in the 10kb region around the TSS. This implies that there is a simpler representation of differential methylation which can be constructed to predict gene expression change. However, this lower resolution representation is still more sufficient for encoding DNA methylation information than a single or even multiple bins. ME-Class is in contrast to first defining the most variable DMR regions and then overlapping these with the largest expression changes (89). Often these studies do not find a strong correlation between methylation and expression, but instead choose the top candidates with an inverse correlation and

discuss those without justification for parameter selection, even though there will be examples of an inverse correlation by random chance. This hypothesis is also supported by single sample DNA methylation methods of deconstructing the methylation curve into a feature representation (189).

5.6 DNA Methylation and Hydroxymethylation

ME-Class demonstrates that 5hmC is insufficient for the prediction of gene expression change apart from a minor class of conversion between 5mC to 5hmC. This small relative change in predictive performance is consistent with the relative scarcity of 5hmC in the genome. This small contribution might also be attributed to a relatively small, measurable change in 5hmC, and continuing development of single cell single base pair resolution 5hmC assays would help to understand the kinetics of active demethylation (190). ME-Class has identified a class 5mC/5hmC patterns that show the conversion from 5mC to 5hmC in the 3' proximal region of the promoter in a model of murine brain development. This class of 5mC conversion, however, is not found in a comparison of human tissue- or cancer-specific 5mC and 5hmC signatures. I hypothesize that these patterns of 5mC and 5hmC are in coordination with specific TF that co-occur with TET enzymes in a context-specific manner. We can take advantage of a proposed model for the relationship between transcription factor (TF) binding affinities and DNA methylation near the TSS (191). We can extend this model to identify these patterns by subsetting TFs that are both methylation- and hydroxymethylation-sensitive and insensitive (192).

References

1. Consortium,R.E., Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J., *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.

2. Adams,D., Altucci,L., Antonarakis,S.E., Ballesteros,J., Beck,S., Bird,A., Bock,C., Boehm,B., Campo,E., Caricasole,A., *et al.* (2012) BLUEPRINT to decode the epigenetic signature written in blood. *Nat Biotech*, **30**, 224–226.

3. Schubeler,D. (2015) Function and information content of DNA methylation. *Nature*, **517**, 321–326.

4. Hotchkiss,R.D. (1948) The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. *J. Biol. Chem.*, **175**, 315–332.

5. Bestor,T.H., Edwards,J.R. and Boulard,M. (2015) Notes on the role of dynamic DNA methylation in mammalian development. *Proceedings of the National Academy of Sciences*, **112**, 6796–6799.

6. Lister,R., Pelizzola,M., Dowen,R.H., Hawkins,R.D., Hon,G., Tonti-Filippini,J., Nery,J.R., Lee,L., Ye,Z., Ngo,Q.-M., *et al.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315–322.

7. Saxonov,S., Berg,P. and Brutlag,D.L. (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences*, **103**, 1412–1417.

8. Bibikova,M., Barnes,B., Tsan,C., Ho,V., Klotzle,B., Le,J.M., Delano,D., Zhang,L., Schroth,G.P., Gunderson,K.L., *et al.* (2011) High density DNA methylation array with single CpG site resolution. *New Genomic Technologies and Applications*, **98**, 288–295.

9. Reik,W. (2001) Epigenetic Reprogramming in Mammalian Development. *Science*, **293**, 1089–1093.

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

11. Guo,H., Hu,B., Yan,L., Yong,J., Wu,Y., Gao,Y., Guo,F., Hou,Y., Fan,X., Dong,J., *et al.* (2016) DNA methylation and chromatin accessibility profiling of mouse and human fetal germ cells. *Cell Res*.

12. Okano,M., Bell,D.W., Haber,D.A. and Li,E. (1999) DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development. *Cell*, **99**, 247–257.

13. Bestor,T.H. (2000) The DNA methyltransferases of mammals. *Human Molecular Genetics*,

9, 2395–2402.

14. Gowher,H., Liebert,K., Hermann,A., Xu,G. and Jeltsch,A. (2005) Mechanism of Stimulation of Catalytic Activity of Dnmt3A and Dnmt3B DNA-(cytosine-C5)-methyltransferases by Dnmt3L. *J. Biol. Chem.*, **280**, 13341–13348.

15. Bird,A., Nan,X., Ng,H.-H., Johnson,C.A., Laherty,C.D., Turner,B.M. and Eisenman,R.N. (1998) Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature*, **393**, 386–389.

16. Chahrour,M., Jung,S.Y., Shaw,C., Zhou,X., Wong,S.T.C., Qin,J. and Zoghbi,H.Y. (2008) MeCP2, a Key Contributor to Neurological Disease, Activates and Represses Transcription. *Science*, **320**, 1224–1229.

17. Hendrich,B. and Bird,A. (1998) Identification and Characterization of a Family of Mammalian Methyl-CpG Binding Proteins. *Molecular and Cellular Biology*, **18**, 6538–6547.

18. Mayle,A., Yang,L., Rodriguez,B., Zhou,T., Chang,E., Curry,C.V., Challen,G.A., Li,W., Wheeler,D., Rebel,V.I., *et al.* (2015) Dnmt3a loss predisposes murine hematopoietic stem cells to malignant transformation. *Blood*, **125**, 629–638.

19. Trowbridge,J.J., Snow,J.W., Kim,J. and Orkin,S.H. (2009) DNA Methyltransferase 1 Is Essential for and Uniquely Regulates Hematopoietic Stem and Progenitor Cells. *Cell Stem Cell*, **5**, 442–449.

20. Broske,A.-M., Vockentanz,L., Kharazi,S., Huska,M.R., Mancini,E., Scheller,M., Kuhl,C., Enns,A., Prinz,M., Jaenisch,R., *et al.* (2009) DNA methylation protects hematopoietic stem cell multipotency from myeloerythroid restriction. *Nat Genet*, **41**, 1207–1215.

21. Li,E., Bestor,T.H. and Jaenisch,R. (1992) Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*, **69**, 915–926.

22. Ito,S., D'Alessio,A.C., Taranova,O.V., Hong,K., Sowers,L.C. and Zhang,Y. (2010) Role of Tet proteins in 5mC to 5hmC conversion, ES cell self-renewal, and ICM specification. *Nature*, **466**, 1129–1133.

23. Pastor,W.A., Aravind,L. and Rao,A. (2013) TETonic shift: biological roles of TET proteins in DNA demethylation and transcription. *Nature reviews. Molecular cell biology*, **14**, 341–356.

24. Branco,M.R., Ficz,G. and Reik,W. (2011) Uncovering the role of 5-hydroxymethylcytosine in the epigenome. *Nat Rev Genet*, **13**, 7–13.

25. Mellen,M., Ayata,P., Dewell,S., Kriaucionis,S. and Heintz,N. (2012) MeCP2 binds to 5hmc enriched within active genes and accessible chromatin in the nervous system. *Cell*, **151**, 1417–1430.

26. Dawlaty,M.M., Breiling,A., Le,T., Raddatz,G., Barrasa,M.I., Cheng,A.W., Gao,Q.,

Powell,B.E., Li,Z., Xu,M., *et al.* (2013) Combined deficiency of Tet1 and Tet2 causes epigenetic abnormalities but is compatible with postnatal development. *Developmental cell*, **24**, 310–323.

27. Dawlaty,M.M., Breiling,A., Le,T., Barrasa,M.I., Raddatz,G., Gao,Q., Powell,B.E., Cheng,A.W., Faull,K.F., Lyko,F., *et al.* (2014) Loss of Tet enzymes compromises proper differentiation of embryonic stem cells. *Developmental cell*, **29**, 102–111.

28. Song,C.-X., Yi,C. and He,C. (2012) Mapping recently identified nucleotide variants in the genome and transcriptome. *Nat Biotech*, **30**, 1107–1116.

29. Eckhardt,F., Lewin,J., Cortese,R., Rakyan,V.K., Attwood,J., Burger,M., Burton,J., Cox,T.V., Davies,R., Down,T.A., *et al.* (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet*, **38**, 1378–1385.

30. Ziller,M.J., Gu,H., Müller,F., Donaghey,J., Tsai,L.T.-Y., Kohlbacher,O., De Jager,P.L., Rosen,E.D., Bennett,D.A., Bernstein,B.E., *et al.* (2013) Charting a dynamic DNA methylation landscape of the human genome. *Nature*, **500**, 477–481.

31. Lehnertz,B., Ueda,Y., Derijck,A.A.H.A., Braunschweig,U., Perez-Burgos,L., Kubicek,S., Chen,T., Li,E., Jenuwein,T. and Peters,A.H.F.M. (2003) Suv39h-Mediated Histone H3 Lysine 9 Methylation Directs DNA Methylation to Major Satellite Repeats at Pericentric Heterochromatin. *Current Biology*, **13**, 1192–1200.

32. Xie,W., Schultz,M.D., Lister,R., Hou,Z., Rajagopal,N., Ray,P., Whitaker,J.W., Tian,S., Hawkins,R.D., Leung,D., *et al.* (2013) Epigenomic Analysis of Multilineage Differentiation of Human Embryonic Stem Cells. *Cell*, **153**, 1134–1148.

33. Berman,B.P., Weisenberger,D.J., Aman,J.F., Hinoue,T., Ramjan,Z., Liu,Y., Noushmehr,H., Lange,C.P.E., van Dijk,C.M., Tollenaar,R.A.E.M., *et al.* (2011) Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina–associated domains. *Nat Genet*, **44**, 40–46.

34. Elliott,G., Hong,C., Xing,X., Zhou,X., Li,D., Coarfa,C., Bell,R.J.A., Maire,C.L., Ligon,K.L., Sigaroudinia,M., *et al.* (2015) Intermediate DNA methylation is a conserved signature of genome regulation. *Nature Communications*, **6**, 6363.

35. Hon,G.C., Rajagopal,N., Shen,Y., McCleary,D.F., Yue,F. and Dang,M.D. (2013) Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. *Nat Genet*, **45**.

36. Edwards,J.R., O'Donnell,A.H., Rollins,R.A., Peckham,H.E., Lee,C., Milekic,M.H., Chanrion,B., Fu,Y., Su,T., Hibshoosh,H., *et al.* (2010) Chromatin and sequence features that define the fine and gross structure of genomic methylation patterns. *Genome Research*, **20**, 972–980.

37. Lorincz,M.C., Dickerson,D.R., Schmitt,M. and Groudine,M. (2004) Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells. *Nat*

Struct Mol Biol, **11**, 1068–1075.

38. Stroud,H., Feng,S., Morey Kinney,S., Pradhan,S. and Jacobsen,S.E. (2011) 5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells. *Genome Biol*, **12**, R54.

39. Lister,R., Mukamel,E.A., Nery,J.R., Urich,M., Puddifoot,C.A., Johnson,N.D., Lucero,J., Huang,Y., Dwork,A.J., Schultz,M.D., *et al.* (2013) Global epigenomic reconfiguration during mammalian brain development. *Science*, **341**, 1237905–1237905.

40. Huang,Y., Chavez,L., Chang,X., Wang,X., Pastor,W.A., Kang,J., Zepeda-Martínez,J.A., Pape,U.J., Jacobsen,S.E., Peters,B., *et al.* (2014) Distinct roles of the methylcytosine oxidases Tet1 and Tet2 in mouse embryonic stem cells. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 1361–1366.

41. Hon,G.C., Song,C.-X., Du,T., Jin,F., Selvaraj,S., Lee,A.Y., Yen,C.-A., Ye,Z., Mao,S.-Q., Wang,B.-A., *et al.* (2014) 5mC Oxidation by Tet2 Modulates Enhancer Activity and Timing of Transcriptome Reprogramming during Differentiation. *Molecular Cell*, **56**, 286–297.

42. Cedar,H. and Bergman,Y. (2009) Linking DNA methylation and histone modification: patterns and paradigms. *Nat Rev Genet*, **10**, 295–304.

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

44. Jenuwein,T. (2001) Translating the Histone Code. *Science*, **293**, 1074–1080.

45. Wang,Z., Zang,C., Rosenfeld,J.A., Schones,D.E., Barski,A., Cuddapah,S., Cui,K., Roh,T.-Y., Peng,W., Zhang,M.Q., *et al.* (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet*, **40**, 897–903.

46. Kouzarides,T. (2007) Chromatin Modifications and Their Function. *Cell*, **128**, 693–705.

47. Bergman,Y. and Cedar,H. (2013) DNA methylation dynamics in health and disease. *Nat Struct Mol Biol*, **20**, 274–281.

48. Voigt,P., Tee,W.-W. and Reinberg,D. (2013) A double take on bivalent promoters. *Genes & Development*, **27**, 1318–1338.

49. Li,E., Beard,C. and Jaenisch,R. (1993) Role for DNA methylation in genomic imprinting. *Nature*, **366**, 362–365.

50. Panning,B. and Jaenisch,R. (1996) DNA hypomethylation can activate Xist expression and silence X-linked genes. *Genes & Development*, **10**, 1991–2002.

51. Walsh,C.P., Chaillet,J.R. and Bestor,T.H. (1998) Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nat Genet*, **20**, 116–117.

52. Riggs,A.D. (1975) X inactivation, differentiation, and DNA methylation. *Cytogenet Genome Res*, **14**, 9–25.

53. Ji,H., Ehrlich,L.I.R., Seita,J., Murakami,P., Doi,A., Lindau,P., Lee,H., Aryee,M.J., Irizarry,R.A., Kim,K., *et al.* (2010) Comprehensive methylome map of lineage commitment from haematopoietic progenitors. *Nature*, **467**, 338–342.

54. Farlik,M., Halbritter,F., Müller,F., Choudry,F.A., Ebert,P., Klughammer,J., Farrow,S., Santoro,A., Ciaurro,V., Mathur,A., *et al.* (2016) DNA Methylation Dynamics of Human Hematopoietic Stem Cell Differentiation. *Cell Stem Cell*, **19**, 808–822.

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

56. Jones,P.A. (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet*, **13**, 484–492.

57. Woodcock,D.M., Lawler,C.B., Linsenmeyer,M.E., Doherty,J.P. and Warren,W.D. (1997) Asymmetric Methylation in the Hypermethylated CpG Promoter Region of the Human L1 Retrotransposon. *J. Biol. Chem.*, **272**, 7810–7816.

58. Liu,W.-M., Maraia,R.J., Rubin,C.M. and Schmid,C.W. (1994) Alu transcripts: cytoplasmic localisation and regulation by DNA methylation. *Nucleic Acids Research*, **22**, 1087–1095.

59. McDonald,J.I., Celik,H., Rois,L.E., Fishberger,G., Fowler,T., Rees,R., Kramer,A., Martens,A., Edwards,J.R. and Challen,G.A. (2016) Reprogrammable CRISPR/Cas9-based system for inducing site-specific DNA methylation. *Biology Open*, **5**, 866–874.

60. Vojta,A., Dobrinić,P., Tadić,V., Bočkor,L., Korać,P., Julg,B., Klasić,M. and Zoldoš,V. (2016) Repurposing the CRISPR-Cas9 system for targeted DNA methylation. *Nucleic Acids Research*, **44**, 5615–5628.

61. Liu,X.S., Wu,H., Ji,X., Stelzer,Y., Wu,X., Czauderna,S., Shu,J., Dadon,D., Young,R.A. and Jaenisch,R. (2016) Editing DNA Methylation in the Mammalian Genome. *Cell*, **167**, 233–247.e17.

62. Cassidy,S.B., Schwartz,S., Miller,J.L. and Driscoll,D.J. (2011) Prader-Willi syndrome. *Genet Med*, **14**, 10–26.

63. Weksberg,R., Shuman,C. and Beckwith,J.B. (2009) Beckwith–Wiedemann syndrome. *Eur J Hum Genet*, **18**, 8–14.

64. Bestor,T.H., Xu,G.-L., Bourc'his,D., Hsieh,C.-L., Tommerup,N., Bugge,M., Hulten,M., Qu,X., Russo,J.J. and Viegas-Péquignot,E. (1999) Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. *Nature*, **402**, 187–191.

65. Zoghbi,H.Y., Amir,R.E., Van den Veyver,I.B., Wan,M., Tran,C.Q. and Francke,U. (1999)

Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nat Genet*, **23**, 185–188.

66. Baylin,S.B. (2005) DNA methylation and gene silencing in cancer. *Nature Clinical Practice Oncology*, **2**, S4–S11.

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

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

69. Irizarry,R.A., Ladd-Acosta,C., Wen,B., Wu,Z., Montano,C., Onyango,P., Cui,H., Gabo,K., Rongione,M., Webster,M., *et al.* (2009) The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat Genet*, **41**, 178–186.

70. Kulis,M., Heath,S., Bibikova,M., Queiros,A.C., Navarro,A., Clot,G., Martinez-Trillos,A., Castellano,G., Brun-Heath,I., Pinyol,M., *et al.* (2012) Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nat Genet*, **44**, 1236–1242.

71. Yoo,C.B. and Jones,P.A. (2006) Epigenetic therapy of cancer: past, present and future. *Nat Rev Drug Discov*, **5**, 37–50.

72. You,J.S. and Jones,P.A. (2012) Cancer Genetics and Epigenetics: Two Sides of the Same Coin? *Cancer Cell*, **22**, 9–20.

73. Yang,H., Liu,Y., Bai,F., Zhang,J.-Y., Ma,S.-H., Liu,J., Xu,Z.-D., Zhu,H.-G., Ling,Z.-Q., Ye,D., *et al.* (2013) Tumor development is associated with decrease of TET gene expression and 5-methylcytosine hydroxylation. *Oncogene*, **32**, 663–669.

74. Stranger,B.E., Forrest,M.S., Dunning,M., Ingle,C.E., Beazley,C., Thorne,N., Redon,R., Bird,C.P., de Grassi,A., Lee,C., *et al.* (2007) Relative Impact of Nucleotide and Copy Number Variation on Gene Expression Phenotypes. *Science*, **315**, 848–853.

75. Mitelman,F., Johansson,B. and Mertens,F. (2007) The impact of translocations and gene fusions on cancer causation. *Nat Rev Cancer*, **7**, 233–245.

76. Beroukhim,R., Mermel,C.H., Porter,D., Wei,G., Raychaudhuri,S., Donovan,J., Barretina,J., Boehm,J.S., Dobson,J., Urashima,M., *et al.* (2010) The landscape of somatic copy-number alteration across human cancers. *Nature*, **463**, 899–905.

77. Kallioniemi,A., Kallioniemi,O., Sudar,D., Rutovitz,D., Gray,J., Waldman,F. and Pinkel,D. (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, **258**, 818–821.

78. Kidd,J.M., Cooper,G.M., Donahue,W.F., Hayden,H.S., Sampas,N., Graves,T., Hansen,N., Teague,B., Alkan,C., Antonacci,F., *et al.* (2008) Mapping and sequencing of structural

variation from eight human genomes. *Nature*, **453**, 56–64.

79. McVean,G.A., Altshuler Co-Chair,D.M., Durbin Co-Chair,R.M., Abecasis,G.R., Bentley,D.R., Chakravarti,A., Clark,A.G., Donnelly,P., Eichler,E.E., Flicek,P., *et al.* (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

80. Feber,A., Guilhamon,P., Lechner,M., Fenton,T., Wilson,G.A., Thirlwell,C., Morris,T.J., Flanagan,A.M., Teschendorff,A.E., Kelly,J.D., *et al.* (2014) Using high-density DNA methylation arrays to profile copy number alterations. *Genome Biol*, **15**, R30.

81. Robinson,M.D., Stirzaker,C., Statham,A.L., Coolen,M.W., Song,J.Z., Nair,S.S., Strbenac,D., Speed,T.P. and Clark,S.J. (2010) Evaluation of affinity-based genome-wide DNA methylation data: Effects of CpG density, amplification bias, and copy number variation. *Genome Research*, **20**, 1719–1729.

82. Bell,J.T., Pai,A.A., Pickrell,J.K., Gaffney,D.J., Pique-Regi,R., Degner,J.F., Gilad,Y. and Pritchard,J.K. (2011) DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. **12**, R10–13.

83. Feng,G., Hobbs,J., Lu,X., Yu,Y., Du,P., Kibbe,W.A., Chandler,J., Hou,L. and Lin,S.M. (2014) A statistical method to estimate DNA copy number from Illumina high-density methylation arrays. *Systems Biomedicine*, **1**, 94–98.

84. Stransky,N., Vallot,C., Reyal,F., Bernard-Pierrot,I., de Medina,S.G.D., Segraves,R., de Rycke,Y., Elvin,P., Cassidy,A., Spraggon,C., *et al.* (2006) Regional copy number–independent deregulation of transcription in cancer. *Nat Genet*, **38**, 1386–1396.

85. Varley,K.E., Gertz,J., Bowling,K.M., Parker,S.L., Reddy,T.E., Pauli-Behn,F., Cross,M.K., Williams,B.A., Stamatoyannopoulos,J.A., Crawford,G.E., *et al.* (2013) Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Research*, **23**, 555–567.

86. Fernandez,A.F., Assenov,Y., Martin-Subero,J.I., Balint,B., Siebert,R., Taniguchi,H., Yamamoto,H., Hidalgo,M., Tan,A.-C., Galm,O., *et al.* (2012) A DNA methylation fingerprint of 1628 human samples. *Genome Research*, **22**, 407–419.

87. Lou,S., Lee,H.-M., Qin,H., Li,J.-W., Gao,Z., Liu,X., Chan,L.L., KL Lam,V., So,W.-Y., Wang,Y., *et al.* (2014) Whole-genome bisulfite sequencing of multiple individuals reveals complementary roles of promoter and gene body methylation in transcriptional regulation. *Genome Biol*, **15**, 6–21.

88. Suzuki,M.M. and Bird,A. (2008) DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet*, **9**, 465–476.

89. Bock,C. (2012) Analysing and interpreting DNA methylation data. *Nat Rev Genet*, **13**, 705–719.

90. Fang,F., Turcan,S., Rimner,A., Kaufman,A., Giri,D., Morris,L.G.T., Shen,R., Seshan,V., Mo,Q., Heguy,A., *et al.* (2011) Breast Cancer Methylomes Establish an Epigenomic

Foundation for Metastasis. *Science Translational Medicine*, **3**, 75ra25–75ra25.

91. Turcan,S., Rohle,D., Goenka,A., Walsh,L.A., Fang,F., Yilmaz,E., Campos,C., Fabius,A.W.M., Lu,C., Ward,P.S., *et al.* (2012) IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. *Nature*, **483**, 479–483.

92. Cabezas-Wallscheid,N., Klimmeck,D., Hansson,J., Lipka,D.B., Reyes,A., Wang,Q., Weichenhan,D., Lier,A., Paleske,von,L., Renders,S., *et al.* (2014) Identification of Regulatory Networks in HSCs and Their Immediate Progeny via Integrated Proteome, Transcriptome, and DNA Methylome Analysis. *Cell Stem Cell*, **15**, 507–522.

93. Wagner,J.R., Busche,S., Ge,B., Kwan,T., Pastinen,T. and Blanchette,M. (2014) The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biol*, **15**, R37–R37.

94. van Eijk,K.R., de Jong,S., Boks,M.P., Langeveld,T., Colas,F., Veldink,J.H., de Kovel,C.G., Janson,E., Strengman,E., Langfelder,P., *et al.* (2012) Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. *BMC Genomics*, **13**, 1–13.

95. Schultz,M.D., He,Y., Whitaker,J.W., Hariharan,M., Mukamel,E.A., Leung,D., Rajagopal,N., Nery,J.R., Urich,M.A., Chen,H., *et al.* (2015) Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature*, **523**, 212–216.

96. Vanderkraats,N.D., Hiken,J.F., Decker,K.F. and Edwards,J.R. (2013) Discovering high-resolution patterns of differential DNA methylation that correlate with gene expression changes. *Nucleic Acids Research*, **41**, 6816–6827.

97. Lund,K., Cole,J.J., VanderKraats,N.D., McBryan,T., Pchelintsev,N.A., Clark,W., Copland,M., Edwards,J.R. and Adams,P.D. (2014) DNMT inhibitors reverse a specific signature of aberrant promoter DNA methylation and associated gene silencing in AML. *Genome Biol*, **15**, 1906–20.

98. Cruickshanks,H.A., McBryan,T., Nelson,D.M., VanderKraats,N.D., Shah,P.P., van Tuyn,J., Rai,T.S., Brock,C., Donahue,G., Dunican,D.S., *et al.* (2013) Senescent cells harbour features of the cancer epigenome. *Nature Cell Biology*, **15**, 1495–1506.

99. DeSantis,C., Ma,J., Bryan,L. and Jemal,A. (2013) Breast cancer statistics, 2013. *CA: A Cancer Journal for Clinicians*, **64**, 52–62.

100. Perou,C.M., Sorlie,T., Eisen,M.B., van de Rijn,M., Jeffrey,S.S., Rees,C.A., Pollack,J.R., Ross,D.T., Johnsen,H., Akslen,L.A., *et al.* (2000) Molecular portraits of human breast tumours. *Nature*, **406**, 747–752.

101. Parker,J.S., Mullins,M., Cheang,M.C., Leung,S., Voduc,D., Vickery,T., Davies,S., Fauron,C., He,X., Hu,Z., *et al.* (2009) Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *Journal of Clinical Oncology*, **27**, 1160–1167.

102. Nielsen,T.O. (2004) Immunohistochemical and Clinical Characterization of the Basal-Like Subtype of Invasive Breast Carcinoma. *Clinical Cancer Research*, **10**, 5367–5374.

103. van 't Veer,L.J., Dai,H., van de Vijver,M.J., He,Y.D., Hart,A.A.M., Mao,M., Peterse,H.L., van der Kooy,K., Marton,M.J., Witteveen,A.T., *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

104. Rouzier,R. (2005) Breast Cancer Molecular Subtypes Respond Differently to Preoperative Chemotherapy. *Clinical Cancer Research*, **11**, 5678–5685.

105. Koboldt,D.C., Fulton,R.S., McLellan,M.D., Schmidt,H., Kalicki-Veizer,J., McMichael,J.F., Fulton,L.L., Dooling,D.J., Ding,L., Mardis,E.R., *et al.* (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.

106. Early Breast Cancer Trialists' Collaborative Group (EBCTCG) (2011) Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials. *The Lancet*, **378**, 771–784.

107. Thomas,C. and Gustafsson,J.-Å. (2011) The different roles of ER subtypes in cancer biology and therapy. *Nat Rev Cancer*, **11**, 597–608.

108. Osborne,C.K. and Schiff,R. (2011) Mechanisms of Endocrine Resistance in Breast Cancer. *Annu. Rev. Med.*, **62**, 233–247.

109. Sanchez,C.G., Ma,C.X., Crowder,R.J., Guintoli,T., Phommaly,C., Gao,F., Lin,L. and Ellis,M.J. (2011) Preclinical modeling of combined phosphatidylinositol-3-kinase inhibition with endocrine therapy for estrogen receptor-positive breast cancer. *Breast Cancer Res*, **13**, R21.

110. Ellis,M.J., Ding,L., Shen,D., Luo,J., Suman,V.J., Wallis,J.W., Van Tine,B.A., Hoog,J., Goiffon,R.J., Goldstein,T.C., *et al.* (2012) Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature*, **486**, 353–360.

111. Edgren,H., Murumagi,A., Kangaspeska,S., Nicorici,D., Hongisto,V., Kleivi,K., Rye,I.H., Nyberg,S., Wolf,M., Borresen-Dale,A.-L., *et al.* (2011) Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome Biol*, **12**, R6.

112. Alkan,C., Coe,B.P. and Eichler,E.E. (2011) Genome structural variation discovery and genotyping. *Nat Rev Genet*, **12**, 363–376.

113. Wood,L.D., Parsons,D.W., Jones,S., Lin,J., Sjoblom,T., Leary,R.J., Shen,D., Boca,S.M., Barber,T., Ptak,J., *et al.* (2007) The Genomic Landscapes of Human Breast and Colorectal Cancers. *Science*, **318**, 1108–1113.

114. Hastings,P.J., Lupski,J.R., Rosenberg,S.M. and Ira,G. (2009) Mechanisms of change in gene copy number. *Nat Rev Genet*, **10**, 551–564.

115. Lupski,J.R. (2007) Genomic rearrangements and sporadic disease. *Nat Genet*, **39**, 543–547.

116. Zack,T.I., Schumacher,S.E., Carter,S.L., Cherniack,A.D., Saksena,G., Tabak,B., Lawrence,M.S., Zhang,C.-Z., Wala,J., Mermel,C.H., *et al.* (2013) Pan-cancer patterns of somatic copy number alteration. *Nat Genet*, **45**, 1134–1140.

117. Aguilar,H., Sole,X., Bonifaci,N., Serra-Musach,J., Islam,A., Lopez-Bigas,N., Mendez-Pertuz,M., Beijersbergen,R.L., Lazaro,C., Urruticoechea,A., *et al.* (2010) Biological reprogramming in acquired resistance to endocrine therapy of breast cancer. *Oncogene*, **29**, 6071–6083.

118. Brown,L.A., Hoog,J., Chin,S.-F., Tao,Y., Zayed,A.A., Chin,K., Teschendorff,A.E., Quackenbush,J.F., Marioni,J.C., Leung,S., *et al.* (2008) ESR1 gene amplification in breast cancer: a common phenomenon? *Nat Genet*, **40**, 806–807.

119. Holst,F., Stahl,P.R., Ruiz,C., Hellwinkel,O., Jehan,Z., Wendland,M., Lebeau,A., Terracciano,L., Al-Kuraya,K., Janicke,F., *et al.* (2007) Estrogen receptor alpha (ESR1) gene amplification is frequent in breast cancer. *Nat Genet*, **39**, 655–660.

120. Moelans,C.B., Monsuur,H.N., de Pinth,J.H., Radersma,R.D., de Weger,R.A. and van Diest,P.J. (2010) ESR1 amplification is rare in breast cancer and is associated with high grade and high proliferation: a multiplex ligation-dependent probe amplification study. *Analytical Cellular Pathology*, **33**, 13–18.

121. Reis-Filho,J.S., Drury,S., Lambros,M.B., Marchio,C., Johnson,N., Natrajan,R., Salter,J., Levey,P., Fletcher,O., Peto,J., *et al.* (2008) ESR1 gene amplification in breast cancer: a common phenomenon? *Nat Genet*, **40**, 809–810.

122. Yang,X., Yan,L. and Davidson,N.E. (2001) DNA methylation in breast cancer. *Endocrine-Related Cancer*, **8**, 115–127.

123. Li,S., Shen,D., Shao,J., Crowder,R., Liu,W., Prat,A., He,X., Liu,S., Hoog,J., Lu,C., *et al.* (2013) Endocrine-Therapy-Resistant ESR1 Variants Revealed by Genomic Characterization of Breast-Cancer-Derived Xenografts. *Cell Reports*, **4**, 1116–1130.

124. Pique-Regi,R., Monso-Varona,J., Ortega,A., Seeger,R.C., Triche,T.J. and Asgharzadeh,S. (2008) Sparse representation and Bayesian detection of genome copy number alterations from microarray data. *Bioinformatics*, **24**, 309–318.

125. Zeitouni,B., Boeva,V., Janoueix-Lerosey,I., Loeillet,S., Legoix-ne,P., Nicolas,A., Delattre,O. and Barillot,E. (2010) SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics*, **26**, 1895–1896.

126. Grabherr,M.G., Haas,B.J., Yassour,M., Levin,J.Z., Thompson,D.A., Amit,I., Adiconis,X., Fan,L., Raychowdhury,R., Zeng,Q., *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotech*, **29**, 644–652.

127. Anders,S., Pyl,P.T. and Huber,W. (2014) HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–169.

128. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2009) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

129. Rozen,S. and Skaletsky,H. (1999) Primer3 on the WWW for General Users and for Biologist Programmers. In *Bioinformatics Methods and Protocols*. Humana Press, pp. 365–386.

130. Hillmer,A.M., Yao,F., Inaki,K., Lee,W.H., Ariyaratne,P.N., Teo,A.S.M., Woo,X.Y., Zhang,Z., Zhao,H., Ukil,L., *et al.* (2011) Comprehensive long-span paired-end-tag mapping reveals characteristic patterns of structural variations in epithelial cancer genomes. *Genome Research*, **21**, 665 –675.

131. Wood,H.M., Belvedere,O., Conway,C., Daly,C., Chalkley,R., Bickerdike,M., McKinley,C., Egan,P., Ross,L., Hayward,B., *et al.* (2010) Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of DNA from formalin-fixed paraffin-embedded specimens. *Nucleic Acids Research*, **38**, e151–e151.

132. Ellis,M.J., Gao,F., Dehdashti,F., Jeffe,D.B., Marcom,P.K., Carey,L.A., Dickler,M.N., Silverman,P., Fleming,G.F., Kommareddy,A., *et al.* (2009) Lower-Dose vs High-Dose Oral Estradiol Therapy of Hormone Receptor–Positive, Aromatase Inhibitor–Resistant Advanced Breast Cancer. *JAMA*, **302**, 774–780.

133. Holst,F. (2016) Estrogen receptor alpha gene amplification in breast cancer: 25 years of debate. *World Journal of Clinical Oncology*, **7**, 160–173.

134. Smith,Z.D. and Meissner,A. (2013) DNA methylation: roles in mammalian development. *Nat Rev Genet*, **14**, 204–220.

135. Laurent,L., Wong,E., Li,G., Huynh,T., Tsirigos,A., Ong,C.T., Low,H.M., Kin Sung,K.W., Rigoutsos,I., Loring,J., *et al.* (2010) Dynamic changes in the human methylome during differentiation. *Genome Research*, **20**, 320–331.

136. Laird,P.W. and Jaenisch,R. (1994) DNA methylation and cancer. *Hum Mol Genet*, **3**, 1487–1495.

137. Baylin,S.B., Esteller,M., Rountree,M.R., Bachman,K.E., Schuebel,K. and Herman,J.G. (2001) Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer. *Human Molecular Genetics*, **10**, 687–692.

138. Ehrlich,M. (2002) DNA methylation in cancer: too much, but also too little. *Oncogene*, **21**, 5400–5413.

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

140. Proudhon,C., Duffié,R., Ajjan,S., Cowley,M., Iranzo,J., Carbajosa,G., Saadeh,H., Holland,M.L., Oakey,R.J., Rakyan,V.K., *et al.* (2012) Protection against De Novo

Methylation Is Instrumental in Maintaining Parent-of-Origin Methylation Inherited from the Gametes. *Molecular Cell*, **47**, 909–920.

141. Yang,X., Han,H., De Carvalho,D.D., Lay,F.D., Jones,P.A. and Liang,G. (2014) Gene Body Methylation Can Alter Gene Expression and Is a Therapeutic Target in Cancer. *Cancer Cell*, **26**, 577–590.

142. Ball,M.P., Li,J.B., Gao,Y., Lee,J.-H., LeProust,E.M., Park,I.-H., Xie,B., Daley,G.Q. and Church,G.M. (2009) Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat Biotech*, **27**, 361–368.

143. Wu,C., MacLeod,I. and Su,A.I. (2013) BioGPS and MyGene.info: organizing online, gene-centric information. *Nucleic Acids Research*, **41**, D561–D565.

144. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,A.D. (2002) The Human Genome Browser at UCSC. *Genome Research*, **12**, 996–1006.

145. Blattler,A., Yao,L., Witt,H., Guo,Y., Nicolet,C.M., Berman,B.P. and Farnham,P.J. (2014) Global loss of DNA methylation uncovers intronic enhancers in genes showing expression changes. *Genome Biol*, **15**, 327–16.

146. Jjingo,D., Conley,A.B., Yi,S.V., Lunyak,V.V. and Jordan,I.K. (2012) On the presence and role of human gene-body DNA methylation. *Oncotarget*, **3**, 462–474.

147. Wu,H., Xu,T., Feng,H., Chen,L., Li,B., Yao,B., Qin,Z., Jin,P. and Conneely,K.N. (2015) Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. *Nucleic Acids Research*, **43**, e141–e141.

148. Hansen,K.D., Langmead,B. and Irizarry,R.A. (2012) BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol*, **13**, R83.

149. Breiman,L. (2001) Random Forests. *Machine Learning*, **45**, 5–32.

150. Pedregosa,F., Varoquaux,G., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Prettenhofer,P., Weiss,R., Dubourg,V., *et al.* (2011) Scikit-learn: Machine Learning in Python. *J Mach Learn Res*, **12**, 2825–2830.

151. Albanese,D., Visintainer,R., Merler,S., Riccadonna,S., Jurman,G. and Furlanello,C. (2012) mlpy: Machine Learning Python. *arXiv*, **1202.6548**, 1–4.

152. Kretzmer,H., Bernhart,S.H., Wang,W., Haake,A., Weniger,M.A., Bergmann,A.K., Betts,M.J., Carrillo-de-Santa-Pau,E., Doose,G., Gutwein,J., *et al.* (2015) DNA methylome analysis in Burkitt and follicular lymphomas identifies differentially methylated regions linked to somatic mutation and transcriptional control. *Nat Genet*, **47**, 1316–1325.

153. Sakoe,H. and Chiba,S. (1978) Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Trans. Acoust., Speech, Signal Process.*, **ASSP-26**, 43–49.

154. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protocols*, **4**, 44–57.

155. Hsieh,C.L. (1994) Dependence of transcriptional repression on CpG methylation density. *Molecular and Cellular Biology*, **14**, 5487–5494.

156. Deaton,A.M. and Bird,A. (2011) CpG islands and the regulation of transcription. *Genes & Development*, **25**, 1010–1022.

157. Bocker,M.T., Hellwig,I., Breiling,A., Eckstein,V., Ho,A.D. and Lyko,F. (2011) Genome-wide promoter DNA methylation dynamics of human hematopoietic progenitor cells during differentiation and aging. *Blood*, **117**, e182–e189.

158. Kulis,M., Merkel,A., Heath,S., Queiros,A.C., Schuyler,R.P., Castellano,G., Beekman,R., Raineri,E., Esteve,A., Clot,G., *et al.* (2015) Whole-genome fingerprint of the DNA methylome during human B cell differentiation. *Nat Genet*, **47**, 746–756.

159. Severson,P.L., Tokar,E.J., Vrba,L., Waalkes,M.P. and Futscher,B.W. (2014) Coordinate H3K9 and DNA methylation silencing of ZNFs in toxicant-induced malignant transformation. *Epigenetics*, **8**, 1080–1088.

160. Herman,J.G. and Baylin,S.B. (2003) Gene Silencing in Cancer in Association with Promoter Hypermethylation. *N Engl J Med*, **349**, 2042–2054.

161. Robertson,K.D. (2005) DNA methylation and human disease. *Nat Rev Genet*, **6**, 597–610.

162. Moarii,M., Boeva,V., Vert,J.-P. and Reyal,F. (2015) Changes in correlation between promoter methylation and gene expression in cancer. *BMC Genomics*, **16**, 142–14.

163. Hovestadt,V., Jones,D.T.W., Picelli,S., Wang,W., Kool,M., Northcott,P.A., Sultan,M., Stachurski,K., Ryzhova,M., Warnatz,H.-J., *et al.* (2014) Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing. *Nature*, **510**, 537–541.

164. Hodges,E., Molaro,A., Santos,Dos,C.O., Thekkat,P., Song,Q., Uren,P.J., Park,J., Butler,J., Rafii,S., McCombie,W.R., *et al.* (2011) Directional DNA Methylation Changes and Complex Intermediate States Accompany Lineage Specificity in the Adult Hematopoietic Compartment. *Molecular Cell*, **44**, 17–28.

165. Azad,N., Zahnow,C.A., Rudin,C.M. and Baylin,S.B. (2013) The future of epigenetic therapy in solid tumours—lessons from the past. *Nature reviews. Clinical oncology*, **10**, 256–266.

166. Nestor,C.E., Ottaviano,R., Reddington,J., Sproul,D., Reinhardt,D., Dunican,D., Katz,E., Dixon,J.M., Harrison,D.J. and Meehan,R.R. (2012) Tissue type is a major modifier of the 5-hydroxymethylcytosine content of human genes. *Genome Research*, **22**, 467–477.

167. Pfeifer,G.P., Kadam,S. and Jin,S.-G. (2013) 5-hydroxymethylcytosine and its potential roles in development and cancer. *Epigenetics & Chromatin*, **6**, 10.

168. Delhommeau,F., Dupont,S., Valle,V.D., James,C., Trannoy,S., Massé,A., Kosmider,O., Le Couedic,J.-P., Robert,F., Alberdi,A., *et al.* (2009) Mutation in TET2in Myeloid Cancers. *N Engl J Med*, **360**, 2289–2301.

169. Chou,W.C., Chou,S.C., Liu,C.Y., Chen,C.Y., Hou,H.A., Kuo,Y.Y., Lee,M.C., Ko,B.S., Tang,J.L., Yao,M., *et al.* (2011) TET2 mutation is an unfavorable prognostic factor in acute myeloid leukemia patients with intermediate-risk cytogenetics. *Blood*, **118**, 3803–3810.

170. Plongthongkum,N., Diep,D.H. and Zhang,K. (2014) Advances in the profiling of DNA modifications: cytosine methylation and beyond. *Nat Rev Genet*, **15**, 647–661.

171. Li,X., Liu,Y., Salz,T., Hansen,K.D. and Feinberg,A.P. (2016) Whole genome analysis of the methylome and hydroxymethylome in normal and malignant lung and liver. *Genome Research*, **26**, gr.211854.116–1741.

172. Schlosberg,C.E., VanderKraats,N.D. and Edwards,J.R. (2017) Modeling complex patterns of differential DNA methylation that associate with gene expression changes. *Nucleic Acids Research*.

173. Qu,J., Zhou,M., Song,Q., Hong,E.E. and Smith,A.D. (2013) MLML: consistent simultaneous estimates of DNA methylation and hydroxymethylation. *Bioinformatics*, **29**, 2645–2646.

174. Liao,Y., Smyth,G.K. and Shi,W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.

175. Timp,W. and Feinberg,A.P. (2013) Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host. *Nature Reviews Cancer*, **13**, 497–510.

176. Montaño,C.M., Irizarry,R.A., Kaufmann,W.E., Talbot,K., Gur,R.E., Feinberg,A.P. and Taub,M.A. (2013) Measuring cell-type specific differential methylation in human brain tissue. *Genome Biol*, **14**, R94.

177. Gravina,S., Dong,X., Yu,B. and Vijg,J. (2016) Single-cell genome-wide bisulfite sequencing uncovers extensive heterogeneity in the mouse liver methylome. *Genome Biol*, **17**, 150.

178. Smallwood,S.A., Lee,H.J., Angermueller,C., Krueger,F., Saadeh,H., Peat,J., Andrews,S.R., Stegle,O., Reik,W. and Kelsey,G. (2014) Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat Meth*, **11**, 817–820.

179. Raine,A., Manlig,E., Wahlberg,P., Syvänen,A.-C. and Nordlund,J. (2016) SPlinted Ligation Adapter Tagging (SPLAT), a novel library preparation method for whole genome bisulphite sequencing. *Nucleic Acids Research*.

180. Lev Maor,G., Yearim,A. and Ast,G. (2015) The alternative role of DNA methylation in splicing regulation. *Trends in Genetics*, **31**, 274–280.

181. Pickrell,J.K., Pai,A.A., Gilad,Y. and Pritchard,J.K. (2010) Noisy Splicing Drives mRNA Isoform Diversity in Human Cells. *PLOS Genetics*, **6**, e1001236.

182. Medvedeva,Y.A., Khamis,A.M., Kulakovskiy,I.V., Ba-Alawi,W., Bhuyan,M.S.I., Kawaji,H., Lassmann,T., Harbers,M., Forrest,A.R. and Bajic,V.B. (2014) Effects of cytosine methylation on transcription factor binding sites. *BMC Genomics*, **15**, 119.

183. Aran,D., Sabato,S. and Hellman,A. (2013) DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. *Genome Biol*, **14**, R21.

184. Shlyueva,D., Stampfel,G. and Stark,A. (2014) Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet*, **15**, 272–286.

185. Visel,A., Minovitsky,S., Dubchak,I. and Pennacchio,L.A. (2007) VISTA Enhancer Browser—a database of tissue-specific human enhancers. *Nucleic Acids Research*, **35**, D88–D92.

186. Smith,Z.D. and Meissner,A. (2013) DNA methylation: roles in mammalian development. *Nat Rev Genet*, **14**, 204–220.

187. Singh,R., Lanchantin,J., Robins,G. and Qi,Y. (2016) DeepChrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics*, **32**, i639–i648.

188. Arvey,A., Agius,P., Noble,W.S. and Leslie,C. (2012) Sequence and chromatin determinants of cell-type–specific transcription factor binding. *Genome Research*, **22**, 1723 –1734.

189. Kapourani,C.-A. and Sanguinetti,G. (2016) Higher order methylation features for clustering and prediction in epigenomic studies. *Bioinformatics*, **32**, i405–i412.

190. Mooijman,D., Dey,S.S., Boisset,J.-C., Crosetto,N. and van Oudenaarden,A. (2016) Single-cell 5hmC sequencing reveals chromosome-wide cell-to-cell variability and enables lineage reconstruction. *Nat Biotech*, **34**, 852–856.

191. Liu,L., Jin,G. and Zhou,X. (2015) Modeling the relationship of epigenetic modifications to transcription factor binding. *Nucleic Acids Research*, **43**, 3873–3885.

192. Maurano,M.T., Wang,H., John,S., Shafer,A., Canfield,T., Lee,K. and Stamatoyannopoulos,J.A. (2015) Role of DNA Methylation in Modulating Transcription Factor Occupancy. *Cell Reports*, **12**, 1184–1195.