Title

Deciphering the Genetic Code of DNA Methylation

Author

Wang, Mengchi

Publication Date

2019

UNIVERSITY OF CALIFORNIA SAN DIEGO

Deciphering the Genetic Code of DNA Methylation

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Mengchi Wang

Committee in charge:

Professor Wei Wang, Chair
Professor Bing Ren, Co-Chair
Professor Joseph Ecker
Professor Trey Ideker
Professor Eran Mukamel

2019

The dissertation of Mengchi Wang is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

____

____

____

____
Co-Chair

____
Chair

University of California San Diego

2019

EPIGRAPH

Wisdom is not a product of schooling but of the lifelong attempt to acquire it.

—Albert Einstein

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my first and foremost gratitude to my advisor, Dr. Wei Wang, who has kindly provided me with outstanding resources and insights throughout my Ph.D. education for the last 6 years. His vision and supervision has been crucial for this thesis and has deeply motivated me to become a better scientist and agile problem-solver.

I’m also indebted to my advisory committee members, Dr. Bing Ren, Dr. Joseph Ecker, Dr. Trey Ideker, and Dr. Eran Mukamel, who have provided me with tremendous support and invaluable advice on both my projects and career development. Special thanks go to Dr. Yin Shen for the experimental collaboration detailed in Chapter 4, as well as the instrumental advice on how to be a better researcher.

In addition, this thesis is not possible without the contribution from my colleagues, Dr. Kai Zhang, Dr. Vu Ngo, Dr. John W Whitaker, Dr. Shicai Fan, Dr. Yue Chen, Dr. Zhao Chen, Dr. Rizi Ai, Dr. Jun Wang, Dr. Guoqiang Li, Dr. Xiaoyu Yang, Dr. Xingjie Ren, Chengyu Liu, Lina Zheng, and David Wang. I thank them for their valuable assistance in the publication included in this thesis.

I would especially like to thank my parents, Zhengping and Shuling, for their unconditional love that transcends time and distance; my wife, Ming, for her unyielding support and beaming encouragement that outshine every hardship along the way; and my dearest friends, Keith and Marguerite, for their cordial friendship and wise guidance that has made San Diego the place that this international student can proudly call home.

Chapter 1, in full, is the material as it would appear as ”Deciphering the Genetic Code of DNA Methylation for Cancer Clinical Application. Mengchi Wang, Vu Ngo, Wei Wang.” Quantitative Biology, 2019. In submission. The dissertation author was a primary investigator and author of this paper.

Chapter 2, in full, is a reprint of the material as it appears in ”Identification of DNA motifs that regulate DNA methylation. Mengchi Wang, Kai Zhang, Vu Ngo, Chengyu Liu, Shicai Fan,

John W Whitaker, Yue Chen, Rizi Ai, Zhao Chen, Jun Wang, Lina Zheng, Wei Wang.” in Nucleic Acid Research, 2019. The dissertation author was a primary investigator and author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in ”Motto: Representing motifs in consensus sequences with minimum information loss. Mengchi Wang, David Wang, Kai Zhang, Vu Ngo, Shicai Fan, Wei Wang.” in Biorxiv, 2019. The dissertation author was a primary investigator and author of this paper.

Chapter 4, in full, is currently being prepared for submission for publication of the material as it would appear as ”CRISPY: a versatile pipeline for CRISPR functional screening. Mengchi Wang, Xiaoyu Yang, Guoqiang Li, Xingjie Ren, Yin Shen, Bing Ren, Wei Wang” The dissertation author was a primary investigator and author of this paper.

VITA

2010	B. S. in Biological Sciences, Nanjing Agricultural University, Nanjing, China
2011-2013	Teaching Assistant, the Ohio State University, Columbus, Ohio
2013	M. S. in Microbiology, the Ohio State University, Columbus, Ohio
2014-2019	Research Assistant, University of California San Diego
2019	Ph. D. in Bioinformatics and Systems Biology, University of California San Diego

PUBLICATIONS

Wang,M, Zhang,K., Ngo,V., Liu,C., Fan,S., Whitaker,J.W., Chen,Y., Ai,R., Chen,Z., Wang,J., et al. Identification of DNA motifs that regulate DNA methylation. *Nucleic Acids Research*, 47, 13, 2019.

Wang,M, Wang,D., Zhang,K., Ngo,V., Fan,S. and Wang,W. Motto: Representing motifs in consensus sequences with minimum information loss. *bioRxiv*, 10.1101/607408, 2019

Wang,M, Ngo,V., Wang,W. Deciphering the Genetic Code of DNA Methylation for Cancer Clinical Application. *Quantitative Biology*, 2019. In submission.

Ngo,V., **Wang,M** and Wang,W. Finding de novo methylated DNA motifs. *Bioinformatics*, 10.1093/bioinformatics/btz079, 2019

Zhang,K., **Wang,M**, Zhao,Y. and Wang,W. Taiji: System-level identification of key transcription factors reveals transcriptional waves in mouse embryonic development. *Sci Adv, 5*, eaav3262, 2019

Fan,S., Huang,K., Ai,R., **Wang,M**, and Wang,W. Predicting CpG methylation levels by integrating Infinium HumanMethylation450 BeadChip array data. *Genomics*, 107, 132-137, 2016

Fan,S., Li,C., Ai,R., **Wang,M**, Firestein,G.S. and Wang,W. Computationally expanding infinium HumanMethylation450 BeadChip array data to reveal distinct DNA methylation patterns of rheumatoid arthritis. *Bioinformatics*, 32, 1773-1778, 2016

Zhu,Y., Chen,Z., Zhang,K., **Wang,M**, Medovoy,D., Whitaker,J.W., Ding,B., Li,N., Zheng,L. and Wang,W. Constructing 3D interaction maps from 1D epigenomes. *Nat. Commun.*, 7, 10812, 2016

Fan,S., Tang,J., Li,N., Zhao,Y., Ai,R., Zhang,K., **Wang,M**, Du,W. and Wang,W. Integrative analysis with expanded DNA methylation data reveals common key regulators and pathways in cancers. *NPJ Genom Med*, 4, 2, 2019

Ngo,V., Chen,Z., Zhang,K., Whitaker,J.W., **Wang,M** and Wang,W. Epigenomic analysis reveals DNA motifs regulating histone modifications in human and mouse. *Proc. Natl. Acad. Sci.*, 10.1073/pnas.1813565116, 2019

Gorkin,D.U., Williams,B.A., Trout,D. and Amrhein,H., **Wang,M**, et al. Systematic mapping of chromatin state landscapes during mouse development. *biorxiv*, 2017

Ai,R., Laragione,T., Hammaker,D., Boyle,D.L., Wildberg,A., Maeshima,K., Palescandolo,E., Krishna,V., Pocalyko,D., Whitaker,J.W., **Wang,M**, et al. Comprehensive epigenetic landscape of rheumatoid arthritis fibroblast-like synoviocytes. *Nat. Commun.*, 9, 1921. 2018

ABSTRACT OF THE DISSERTATION

Deciphering the Genetic Code of DNA Methylation

by

Mengchi Wang

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2019

Professor Wei Wang, Chair
Professor Bing Ren, Co-Chair

DNA methylation plays crucial roles in many biological processes and abnormal DNA methylation patterns are often observed in diseases. Recent studies have shed light on cis-acting DNA elements that regulate locus-specific DNA methylation. More importantly, these new discoveries have shown potentials in clinical application.

In this thesis, I first interrogate the current biological foundation for the cis-acting genetic code that regulates DNA methylation. This process involves transcription factors, histone modifications, and DNA secondary structure. In chapter 2, we demonstrate how to find the functional motifs that regulate DNA methylation. We have analyzed 34 diverse whole-genome bisulfite sequencing datasets and have identified 313 identified motifs, including 92 and 221 associated with methylation (methylation motifs, MMs) and unmethylation (unmethylation motifs, UMs), respectively. We show that these motifs are associated with local methylation level, and motif disruption of by mutation leads to significantly altered methylation level of the CpGs in the neighbor regions. Combined with somatic mutations, these motifs improve the prediction of cancer subtypes and patient survival.

DNA motif analysis frequently requires intuitive understanding and convenient representation of motifs. In chapter 3, I review how the motifs are typically represented as position weight matrices (PWMs) and propose a new wildcard-style consensus sequence representation based on mutual information theory and Jenson-Shannon Divergence. We name this representation as sequence Motto and have implemented an efficient algorithm with flexible options for converting motif PWMs into Motto from nucleotides, amino acids, and customized alphabets. On the other hand, experimental validation of cis-acting DNA elements benefits from the recent advancement of CRISPR/Cas9 mediated genetic screening. In chapter 4, I present CRISPY, a lightweight, robust CRISPR screening pipeline that unifies single-sgRNA and CREST-seq screening protocols and is capable of profiling peak candidates with existing data of histone modifications, DHS, and ATAC-seq in human and mouse.

Combined together, our studies have provided new insights on how genetic code regulates DNA methylation and can be applied to clinical applications. In addition, we provide the tools to efficiently represent the motifs and evaluate their functions in a high-throughput manner.

Chapter 1

The biological foundation for genetic code that regulates methylation

1.1 INTRODUCTION

DNA methylation is the addition of a methyl group on cytosines predominantly at CpG dinucleotides, which could alter how proteins bind to this region. Where DNA is methylated leads to different biological functions. For example, hypermethylation in promoters plays notable roles in gene silencing, whereas DNA methylation at the gene body is involved in transcription elongation and alternative splicing[1, 2]. DNA methylation also synergizes with local histone modification during development, somatic cell reprogramming, and tumorigenesis[3, 4]. Therefore, to interpret the function of DNA methylation, we need to understand how DNA methylation takes place in a locus-specific manner.

Locus-specific DNA methylation or demethylation depends on the recruitment of specific enzymes such as TET and DNMTs to particular genomic regions[5–7]. DNA methylation is catalyzed by DNA methyltransferases (DNMTs) by two distinct mechanisms[8]. New methylation is established (or de novo DNA methylation) on both DNA strands by DNMT3A/3B/3L, predominantly during embryogenesis. Meanwhile, existing DNA methylation is maintained by DNMT1 (with help from UHRF1), which recognizes half-methylated DNA strand (hemimethylation) after replication. On the other hand, the removal of the methyl group from cytosines is promoted by the ten-eleven-translocation enzymes (TET1/2/3), which can oxidize 5mC to 5-hydroxymethylcytosine, and then demethylate to cytosine through various pathways[9].

Emerging evidence has suggested that enzymes like DNMTs and TETs are recruited to specific genomic regions by certain DNA sequences. Recently, we have published a study that systematically identified 313 DNA motifs that regulates DNA methylation from 34 whole-genome methylomes. We show that these motifs are functional and can be applied to improve cancer prognosis and diagnosis[10]. In this review, we aim to first summarize the underlying mechanisms where cis-acting genetic code mediates DNA methylation. Further, we review the trend of recent machine learning models that focus on genetic features of DNA methylation and discuss the biological insights revealed from these models. Finally, we propose to combine DNA methylation genetic code and DNA variants in clinical settings, particularly in liquid biopsy and early cancer diagnosis. We show how this improves the current paradigm where the discovery of biomarkers has been focused on a handful of genes.

1.2 THE MECHANISMS OF LOCUS-SPECIFIC METHYLATION

Overwhelming evidence has shown that DNA methylation exhibit certain genetic patterns, such as lower GC content, enrichment of short nucleotide combinations (2-6 bp), and longer DNA motifs[11–21]. Some experiments further report cases that some DNA sequences can dictate where DNA methylation/demethylation takes place. For example, Lienert *et al* have identified methylation-determining regions, which mediates *de novo* methylation and demethylated. Interestingly, in these regions there are *cis-regulatory* motifs that can be recognized by DNA-binding factors (SP1, CTCF, Rfx); mutating these motifs nullifies the methylation alteration[22]. Stadler et al. have shown the introducing the CTCF motif site are necessary and sufficient to lower methylation of local CpGs[23]. Taken together, these reports suggest methylation is encoded genetically, recognized and mediated by locus-specific protein factors. Here, we review the emerging mechanisms of locus-specific DNA methylation guided by *cis*-acting DNA sequences, through the assistance of a variety of mechanisms, including transcription factors, DNMTs, TETs, DNA secondary structures, and crosstalk from histone modifications (**Figure 1.1**).

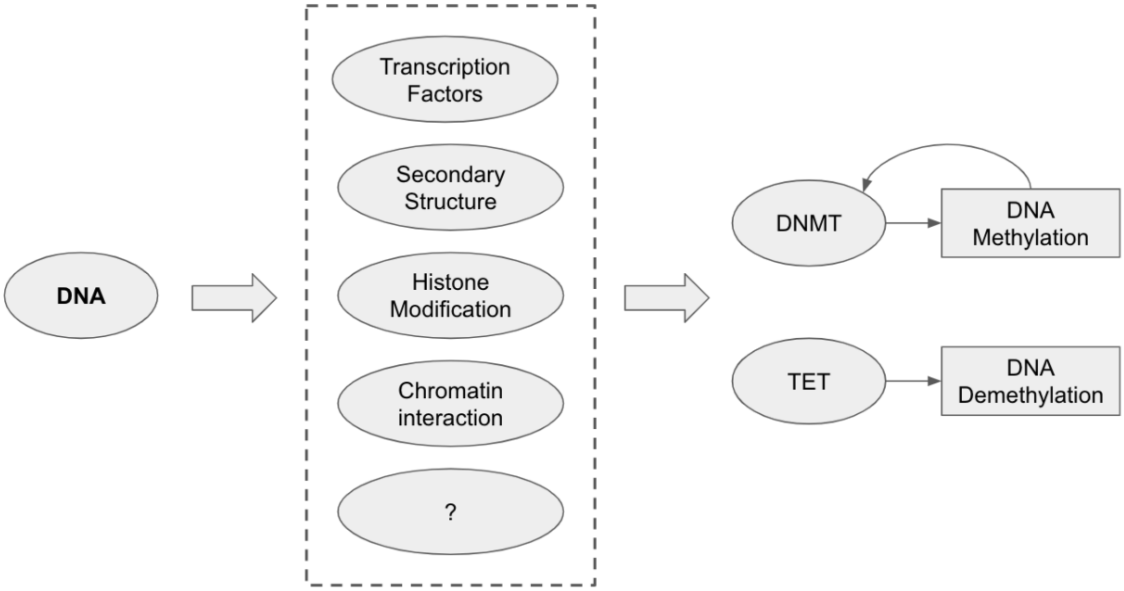


Figure 1.1: The emerging view of locus-specific DNA methylation and demethylation.

1.2.1 TFs recruit TETs for active demethylation.

Previous reports show TET prefers CpG-rich patterns such as CpG island which spans several kilobases[24] and can bind CpG-rich DNA sequences[5] in mammalians to maintain stable demethylation[25]. TET recruitment through locus-specific TF binding has been abundantly reported. For example, introducing a CTCF binding site at a particular locus leads to TET recruitment and local DNA demethylation[23]. PPARG binds to the promoter containing its binding sequence and recruits TET, resulting in local DNA demethylation[26]. In a more recent study, Suzuki *et al.* have designed a method to screen for TFs that can facilitate the demethylation of DNA in a site-directed manner. In particular, they transduced the target TF under test in sub-cloned vectors to target cells, and then test the change in methylation status of the CpGs (by using the HumanMethylation450 methylation array) nearby the TF binding sites (by searching for known motifs on the genome) with and without the ectopic expression of the TF, and screened for methylation change. Using this strategy, Suzuki *et al.* have shown RUNX1 site-specific binding correlates with demethylation in hematopoietic cells, and have further confirmed recruitment of DNA demethylation machinery enzymes including TET2, TET3, TDG, and GADD45, using co-immunoprecipitation[27]. In a separate study, Suzuki *et al.* scaled-up the same strategy and confirmed 8 (RUNX3, GATA2, CEBPB, MAFB, NR4A2, MYOD1, CEBPA, and TBX5) out of 15 (plus NANOG, HNF1A, PAX4, Nkx2-5, SOX2, POU5F1, HNF4A) tested TFs can facilitate demethylation of DNA in a site-directed manner [28].

1.2.2 TFs blocks DNMT3s and prevent *de novo*methylation

Many transcription factors (TFs) can maintain a low methylation region through blocking DNMTs. For example, SP1 preferentially binds to CpG-rich promoters, blocking the region from *de novo* methylation in mouse[29, 30]. Proteins containing a CXXC domain (CFP1, MLL, KDM2A/2B, IDAX) can bind to unmethylated CpGs to prevent the region from methylation[31–33]. Interestingly, DNMT1 has a CXXC domain, which putatively helps it to bind to hemimethylated CpGs[34]; TET1 and TET3 also have a CXXC domain, which has been indicated to contribute to locus-specific genomic recruitment[35]. However, other studies have shown that the CXXC domain failed to restrain the activity of Dnmt1 on unmethylated CpG sites [36].

1.2.3 TFs recruits DNMTs for *de novo* methylation

Similarly, many TFs have been reported to facilitate DNA methylation through locus-specific interaction with their binding sites. For example, NR6A1(or GCNF) can silence Oct-3/4 by binding to its promoter and recruit Dnmt3a and Dnmt3b in the mouse, facilitating methylation[37]. Dnmt3a has been reported to interact with Myc and specifically target the promoter of p21Cip1, leading to repressed transcription[38]. Dnmt3b has been reported to be recruited through E2F6 transcriptional repressor leading to germ-line gene silencing in murine somatic tissues[39].

1.2.4 DNA secondary structure shape DNA methylation

Besides TF-directed locus-specific methylation, DNA secondary structure has also been reported to shape local DNA-methylation. For example, Clark and Smith showed that VNTR (variable number tandem repeats) at a non-B DNA structure contributes to the abnormal DNA methylation in human breast cancers[40]. Mao et al. report G-quadruplex (G4) DNA secondary structures are associated with hypomethylation at the CpG island in the human genome. Paradoxically, G4 sites are enriched with DNMT1 binding, but inhibits DNMT1 enzymatic activity, leading to the inhibition of local CpG methylation [41]. Other studies show a certain group of G4 structures play roles in both DNA methylation and histone modification[42]. Meanwhile, G4 secondary structures are characterized by strong telomeric repeats, with cis-acting DNA motifs such as (GGGGCC)(n), TG(4)T(2)G(4)T, and GGGCT(4)GGGC[43–45]. Taken together, the DNA secondary structure adds another layer for how DNA sequence maintains and alter local methylation.

1.2.5 Compound mechanisms

Some aspects of DNA methylation is complicated and involves multiple modes of action from the same factor. For example, reports have shown SPI1 can mediate both *de novo* methylation (by interacting with DNMT3B) and demethylation (by interacting with TET2) in a site-specific manner[46, 47]. The aforementioned CTCF is another complicated example that employs multiple modes of action. Some studies show CTCF can promote unmethylation through blocking DNMTs. For example, Schoenherr *et al.* showed that mutating CTCF-binding sites resulted in the recruitment of DNMTs, leading to increased methylation at the imprinting control region of Igf2/H19 locus in mouse[48]. However, Stadler *et al.* showed that CTCF binding to target motif sites actively creates a low methylation region through the presence of TETs[23]. Other studies showed that CTCF facilitates histone modification and open chromatin, although the causality in relation to DNA methylation remains unclear[49–51].

1.2.6 Crosstalk to histone modification

The maintenance of DNA methylation also involves intricate crosstalk with histone modification. For example, studies have established DNA maintenance on Uhrf1, which recruits Dnmt1 and is essential for ubiquitination of histone H3 at lysine 23 at DNA replication sites, converting hemimethylated DNA to fully methylated DNA[52]. DNA methylation has also been linked to H3K9me3 and H3K27me3, where the H3K9 methyltransferase SETDB1 interacts with DNMT3A and 3B[53, 54]. Interestingly, SETDB1 is not by itself DNA-binding, but form a repression complex with TRIM28 and zinc finger such as ZNF274 to achieve locus-specificity[54, 55]. Viré et al. showed that the H3K27 methyltransferase EZH2, a component of the polycomb repressive complex PRC2, can interact with DNMT1, DNMT3A, and DNMT3B. EZH2[56]. A more recent study by Baubec et al. using genome-wide ChIP-seq and methylome confirmed that DNMT3A and DNMT3B are localized to methylated CpG-dense regions in mouse stem cells. Notably, they found that SETD2-mediated H3K36 methylation interacts with the PWWP domain of DNMT3B, leading to DNMT3B preferentially binds and methylate the body of actively transcribed genes[57].

On the other hand, DNA unmethylation can be mediated by TF-mediated co-repression through H3K4 methylation. Cfp1 has been reported to box recruit of H3K4 methyltransferases to promote H3K4me3, help to prevent local CpG island from methylation in mouse embryonic stem cellsbox. However, Cfp1 knockout is insufficient to remove local hypomethylation, suggesting other factors are involved in this process[31, 58]. In another report, unmethylated H3K4 tails have been shown to interact with de novo methylation machinery, such as Dnmt3L and Dnmt3a in mouse[59]. The association between H3K4 methylation and allele-specific DNA methylation has been shown at imprinted loci as well[60], with guidance from TFs like KDM1B[61].

Combined together, these reports depict a comprehensive landscape where the genetic code underlies the locus-specific DNA methylation through various mechanisms and machinery.

1.3 THE MODELS: PREDICTION AND REVELATION

The molecular mechanisms described above have laid the foundation for many studies that use genetic features to predict local DNA methylation. These studies have shed light on the sequence features of locus-specific methylation and demethylation. Below we review the development of these studies and evaluate the trend.

Earlier methylation studies typically employ enzymatic fractionation assays. For example, McrBC digests methylated sequences while many methylation-sensitive restriction endonucleases remove unmethylated sequences[62]. Due to the limited data coverage and resolution, these studies tend to focus on the methylation status of CpG islands (CGI). These CpG islands reside at the promoter of genes and serve important roles during transcription by allowing transcription factors if unmethylated[63]. To distinguish unmethylated CpG-island (non-CGI) from methylated

CpG island (CGI), a variety of predictive features have been found with machine learning models. For example, Yamada et al.[17] showed CG, CT, and CA are the most predictive dinucleotides features for human CpG island states. Das et al.[13] showed that Alu coverage and the certain hexamers are the most predictive (with 86% accuracy) among $\sim$ 100 predefined features (CG content, dinucleotide counts, trinucleotide counts, etc.). Performance is further improved when including non-sequence features such as trinucleotide physicochemical properties[14] (i.e., bendability, nucleosome-rigid, and nucleosome-positioning), histone modification[16, 64], and the methylation states of flanking CpGs[20].

Recent studies take advantage of genome-wide methylation assays, such as 450K array, RRBS, and WGBS. The expanded coverage of methylomes has profoundly changed the locus-specific analysis of DNA methylation in several ways. For example, functional motifs have been found outside of CpG islands, extending into non-coding regions[10, 21]. In addition, genomic and epigenetic data from multiple cell lines and tissues have been made available by consortium efforts such as ENCODE[65], ROADMAP[66], TCGA[67], and iHEC[68]. Methylation levels are compared across multiple tissues, cell lines, and species to establish variability. For example, Zeng et al [21] have analyzed 50 RRBS + 1 WGBS datasets and established the impact of DNA variant on local methylation. Wang et al[10] have identified genomic regions and motifs associated with common and variable methylation across 34 WGBS, validated in 32 450K arrays. More datasets have also allowed more sophisticated machine learning models, such as neural network[15, 19, 21], to outperform previously best-performing machine learning models like SVM and random forests[13, 17, 20, 64]. DNA sequence features have shifted from using predefined sequences and short kmer combinations (usually 2-5 bp)[13, 16, 17, 20, 64] to using longer *de novo* motifs ($> 9bp$)[10, 11, 15, 21, 69, 70]. These studies revealed novel perspectives on how certain genetic patterns can play more role in regulating DNA methylation.

However, the most fundamental change in methylation motif studies is from making predictions to providing biological insights using DNA motifs. For example, many studies of discovered *de novo* motifs have been matched to known TFs to provide a reference of their functions[10, 11, 15, 21, 69, 70]. As a result, while earlier studies have associated hypomethylation with high GC contents[16, 17], recent studies have further provided an explanation with the contributing DNA motifs with repeating GC tandems that are matched to known TFs associated with TET recruitment, such as CTCF, SP family, and WT1[10, 21]. On the other hand, contrary to the previous believe that methylated regions have aberrant transcription factor binding, some TFs have also been found to preferentially binds to highly methylated regions. For example, Xuan-lin et al[70] cross-referenced ChIP-seq of TFs and WGBS to identify over 500 TFBS termed as MethMotifs. Ngo et al[69] proposed a high throughput pipeline that can find methylated motif and discovered motifs that have ”dual-modes” where DNA methylation can act in a sequence-specific context in gene regulations. Whitaker et al. have further provided a computational framework to identify DNA motifs representing cis-acting elements with the site-specific DNA-binding factors that establish and maintain epigenomic modifications[11].

Aside from known TFs, the function of the identified motifs can also be validated by DNA variants. For example, Wang et al.[10] have shown motifs with enriched mQTL and eQTL, and somatic mutation on the motifs correlates with altered local CpG methylation. Similarly, Zeng et al [21] have proposed a deep learning framework, CpGenie, to characterize methylation change from sequence variant, given the neighboring methylation and DNA sequences. Finally, recent studies have highlighted crosstalk between DNA methylation and histone modification among these motifs, especially between H3K36me3 and methylation motifs, as well as between H3K27ac/H3K4me3 and unmethylation motifs[10, 71].

Taken together, we have observed explosive growth of computational models that explain DNA methylation based on sequence features, in combination with the traditional usage of physicochemical properties, nearby CpG states, TF occupancies, and histone states. The improved model performance and revealed genetic-epigenetic association has made clinical application possible.

1.4 CLINICAL APPLICATION

DNA methylation is closely linked to development, aging, and cancer[72, 73]. Here, we evaluate the theoretical basis and preliminary evidence for how to take advantage of these newly unearthed genetic rules for epigenetic modification. In particular, we discuss the potential applications in light of the recently popular liquid biopsy for cancer diagnosis and treatment guidance.

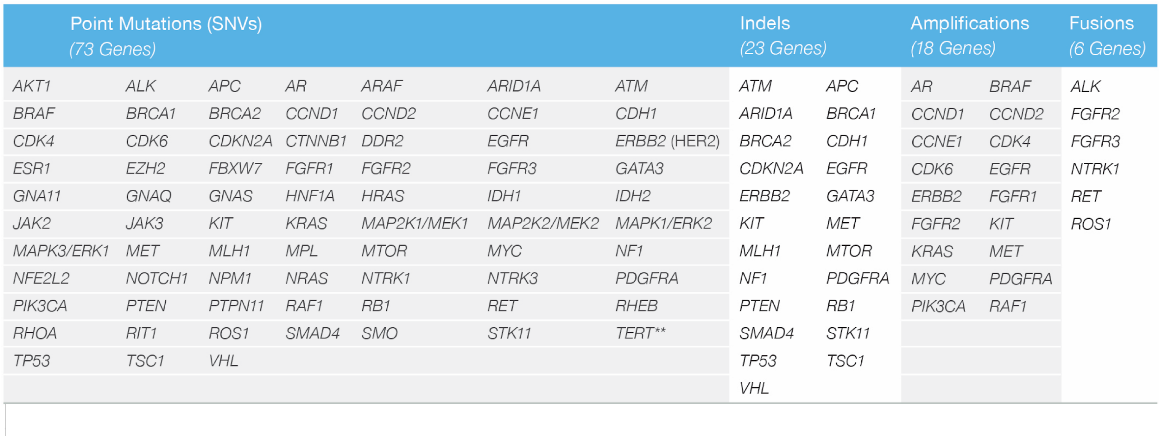
Recently, liquid biopsy has gained great attention and success in early cancer diagnosis and prognosis[74, 75]. When tumor cells go through apoptosis, ctDNA (circulating tumor DNA) is released into the plasma. Along with other DNA fragments found in the plasma, they are collectively termed as ”cfDNA”, or cell-free DNA[74]. Compared to traditional on-site tissue biopsy, liquid biopsy has several distinct advantages. First, it enables minimally-invasive sampling from the blood, or other bodily fluids such as saliva, urine, and stool. Because this sampling is free from the knowledge of target tissue, which happens late in cancer development, liquid biopsy enables the detection of ctDNA in early pathological stages. Finally, given the short half-life ($< 1hr$) of cfDNA, the monitoring is considered ”real-time” , very responsive to environmental changes and cancer treatment[74, 75].

Successful application of liquid biopsy depends on differentiating the tumorous ctDNA from the ”normal” cfDNA fragments. The major challenge is that ctDNA is a small fraction ($0.01\% - 10\%$) of the total cfDNA[74–76]. Therefore, given the finite sequencing power, tumor variant detection relies heavily on knowing how to choose the most predictive set of predictors (or biomarkers), and how to link them to target phenotype. While early research took advantage of simple traits like the level and fragment length (ctDNA $\sim 140bp$, cfDNA $\sim 165bp$) in certain cancer types, recent success came from researching biomarkers to provide more sensitive and selective assay in cancer diagnosis and personalized treatment[75, 77–79].

The traditional biomarkers for cancer are DNA-based, which are usually discovered through cancer-related genetic mutation or variant, including point mutations, copy number alterations, and structural variation[75]. For example, Guardant Health has provided a DNA-based cfDNA assay (Guardant360) to provide better treatment outcome for patients with advanced cancers (such as non-small cell lung cancer). They report sequencing variant detection down to $0.02\% - 0.04\%$ allelic fraction with $> 98\%$ specificity. This is achieved with the help of SNVs from 73 genes, indels from 23 genes, amplifications from 18 genes, and fusions from 6 genes (**Figure 1.2**). These are usually widely reported cancer gatekeeper genes, including KRAS, TP53, BRCA1/2, and BRAF [80].

Compared to late-stage cancer improvement, early pan-cancer detection is a more desirable goal and has been the latest focus for many academic groups combining forces with industry leaders, such as Guardant Health, Grail, Freenome, and CancerSEEK. However, this task is daunting due to two major challenges: (1) early cancer have lower variant level than in late-stage patients; (2) the variant origin is unknown. Current strategies for improving variant detection and locating tissue-of-origin involves combining DNA variants with additional markers. For example, Snyder et al. have reported cfDNA nucleosome occupancies mapped by transcription factor sequence motifs correlate well with the nuclear architecture and cell-specific gene expression, which could be used to inform the tissue-of-origin[81]. Cohen et al. have developed CancerSEEK, which combines DNA variant with protein biomarkers and have reported both detection and location of multiple early-stage cancers, with a sensitivity of 69 to 98% (depending on cancer type) and 99% specificity[82].

Recently, more groups leverage information from the methyl-CpG on cfDNA and exploit the strong link between methylation and cancer[72, 83]. This is based on the observation that methylation on the promoter of a tumor suppressor gene allele can result in a similar function as a genetic alteration[1, 84]. For example, BRCA1, PTEN, HRK, APC, and RASSF1A have been found methylated in cancer, and some related to prognosis and reflect on the efficacy of therapy[85–87]. Henriksen et al. have reported a panel of 28 hypermethylation sites for ctDNA



Point Mutations (SNVs)							Indels		Amplifications		Fusions
(73 Genes)							(23 Genes)		(18 Genes)		(6 Genes)
AKT1	ALK	APC	AR	ARAF	ARID1A	ATM	ATM	APC	AR	BRAF	ALK
BRAF	BRCA1	BRCA2	CCND1	CCND2	CCNE1	CDH1	ARID1A	BRCA1	CCND1	CCND2	FGFR2
CDK4	CDK6	CDKN2A	CTNNB1	DDR2	EGFR	ERBB2 (HER2)	BRCA2	CDH1	CCNE1	CDK4	FGFR3
ESR1	EZH2	FBXW7	FGFR1	FGFR2	FGFR3	GATA3	CDKN2A	EGFR	CDK6	EGFR	NTRK1
GNA11	GNAQ	GNAS	HNF1A	HRAS	IDH1	IDH2	ERBB2	GATA3	ERBB2	FGFR1	RET
JAK2	JAK3	KIT	KRAS	MAP2K1/MEK1	MAP2K2/MEK2	MAPK1/ERK2	KIT	MET	FGFR2	KIT	ROS1
MAPK3/ERK1	MET	MLH1	MPL	MTOR	MYC	NF1	MLH1	MTOR	KRAS	MET	
NFE2L2	NOTCH1	NPM1	NRAS	NTRK1	NTRK3	PDGFRA	NF1	PDGFRA	MYC	PDGFRA	
PIK3CA	PTEN	PTPN11	RAF1	RB1	RET	RHEB	PTEN	RB1	PIK3CA	RAF1	
RHOA	RIT1	ROS1	SMAD4	SMO	STK11	TERT**	SMAD4	STK11			
TP53	TSC1	VHL					TP53	TSC1			
							VHL				

Figure 1.2: Example DNA-based biomarkers from Guardant360 panel reprinted from [80]

as prognostic markers for pancreatic adenocarcinoma staging[88]. DNA methylation patterns derived from RRBS have also been used as a predictor for breast cancer dissemination[89]. Other studies have reported success with DNA methylation cfDNA assay outside plasma for specific cancer types, such as urine-based assays for prostate cancer[90, 91] and stool-based assays for colorectal cancers[92]. Guo et al. reported segments of DNA methylation (termed haplotype blocks) from plasma DNA can aid the deconvolution of heterogeneous tissue samples[93]. A more recent study by Grail has successfully mapped and identifying tumor origin by cfDNA methylation in 25 human tissues and cells[94]. Notably, Shen et al [95] have developed an immunoprecipitation based genome-wide cfDNA methylome screening protocol (cfMeDIP–seq). They showed sensitive tumor detection and classification among several tumor types, using differentially methylated regions and CpGs. Overall, current adoption of methylation in cfDNA focus on either individual genes (particularly at the promoter regions) or differentially methylated regions, and show improvement in otherwise low-performing cancer types using DNA-only biomarkers.

Ultimately, given the finite sequencing power and detection limit, the focus is on how to extract the most phenotypic information from given variants. Recent studies have provided a new strategy, where mutation on *cis-acting* DNA elements leads to altered local methylation and phenotypic association to cancer. For example, an SNP at the MGMT risks the promoter methylation in glioblastoma and is predictive of cancer treatment using temozolomide[96]. An SNP at the CpG site located at the ARPC3 promoter is associated with hypertriglyceridemia in overweight patients[97]. Three CpG-SNP pair has been reported significant for the prognosis of breast cancer patients [98]. Multiple studies have reported DNA variants are particularly found in the CpG island at the promoter of genes related to cancer[99–102]. Zeng et al[21] have reported a model to accurately quantified how DNA variants can impact local CpG methylation and gene expression. Recently, we have discovered and characterized 313 DNA motifs that regulate DNA methylation and unmethylation, and showed that DNA mutation overlapping with these motifs impact local CpG methylation. Moreover, we have demonstrated that profiling somatic mutations in cancer patients based on which DNA motifs they overlap, providing a significant performance improvement over using these somatic mutations alone, both for diagnosis and prognosis [10]. Combined together, these results suggest understanding how DNA-variants change methylation can improve the re-evaluation of the existing DNA biomarkers, and provide new perspectives on biomarker discovery.

Chapter 1, in full, is the material as it would appear as ”Deciphering the Genetic Code of DNA Methylation for Cancer Clinical Application. Mengchi Wang, Vu Ngo, Wei Wang.” Quantitative Biology, 2019. In submission. The dissertation author was a primary investigator and author of this paper.

Chapter 2

Identification of DNA motifs that regulate DNA methylation

2.1 INTRODUCTION

DNA methylation plays crucial roles in many biological processes and aberrant DNA methylation patterns are often observed in diseases. There are three DNA methyltransferases (DNMTs) in human that are responsible for *de novo* or maintaining methylation of cytosine. Although these enzymes themselves do not show strong sequence preference *in vivo*, DNA methylation is highly locus-specific such as hypo-methylation of active promoters and enhancers. An urging question is how such a locus-specific DNA methylation pattern is established. One of the possible mechanisms is that DNA binding proteins or non-coding RNAs recognize specific DNA motifs and their binding recruits DNMTs to a particular locus to methylate cytosines in the region. These factors can be specifically active in a cell type or state such that to provide the cell type- and locus-specificity. Accumulating evidence suggests that protein binding such as CTCF and other proteins can create low methylated regions in the regulatory sites and introducing specific nucleotide sequences can establish DNA methylation[22, 23]. These observations suggested the importance of the DNA sequence in shaping the methylation state. Several studies have illustrated the relationship between sequence features and DNA methylation[11–21],but the DNA motifs recognized by the DNA methylation associated proteins have not been well characterized. Therefore, cataloging these motifs would pave the way towards understanding the mechanism of the locus-specificity of DNA methylation.

Cataloging DNA methylation associated motifs requires a comprehensive set of methylomes and whole-genome bisulfite sequencing (WGBS) is a common technology to map DNA methylation in the entire human genome. The NIH Roadmap Epigenomics Project[103] has generated WGBS data in 34 cell lines or tissues, which provides an opportunity to discern motifs associated with DNA methylation. We reasoned that contrasting regions that are commonly methylated across cells/tissues to those commonly unmethylated would increase the signal-to-noise ratio to identify the motifs most relevant to DNA methylation. Furthermore, to consider the impact of cell type and cell state on DNA methylation, we also need to uncover motifs associated with variable methylation levels across cells/tissues; a caveat is that these motifs can be confounded by those only related to cell specificity. To this end, we have defined commonly methylated (unmethylated) regions across the 34 cells (CMR/CUR) as well as variably methylated (unmethylated) regions (VMR/VUR) that show cell-specific methylation (unmethylation). We have found the DNA motifs that are discriminative of these regions.

To confirm the association with methylation, we overlapped the motifs with DNMT and TET ChIP-seq peaks and observed strong enrichment. We also used TCGA (The Cancer Genome Atlas) dataset to further assess the importance of these motifs in shaping DNA methylation. Interestingly, we found that, if these are somatic mutations occurring in the motifs, the methylation levels in the nearby CpGs are significantly altered, i.e. perturbation to a MM (UM) motif in a highly (lowly) methylated region would decrease (increase) the local methylation level. This observation strongly supports the functionality of the identified motifs in establishing or maintaining locus-specific DNA methylation. Furthermore, we observed eQTLs (expression quantitative trait loci) and mQTLs (methylation quantitative trait loci) are enriched in the found motifs. We also found that the combination of somatic mutations and the found motifs can significantly improve the prediction accuracy of cancer type and patient survival than using somatic mutations alone. This observation also supported the functionality of the DNA methylation associated motifs. Additional analyses also revealed the potential interplay between DNA methylation and histone modification as well as their contribution to DNA methylation dynamics.

2.2 MATERIAL AND METHODS

2.2.1 De novo motif discovery

11.5 million CpG sites common across all human 34 methylomes have been collected from the NIH Roadmap Epigenomics Project[66]. Methylation regions are defined by segments merged with 2 or more CpGs with a maximal distance of 400 bp apart (i.e. CpGs and only CpGs within 400bp of each other will be merged into a methylation region) and region methylation level is defined by the mean CpG beta values. Each region is then assigned mean and standard deviation of methylation across all 34 tissues and cells. we used a normalized score to measure the overall methylation level of a methylation region across 34 methylomes in comparison to the whole genome methylation distribution:

$$score = \frac{\mu_r - \mu_g}{s_r}$$

where μ_r and s_r are the mean and standard deviation of the methylation of the region, μ_g is the mean of methylation genome-wide. We used the ranking of this score in our analysis to select the methylation regions, i.e., CMRs are the CpGs with the top 0.5% score and CURs bottom 0.5% score, while VMRs are defined by the top 20% standard deviation (**Figure 2.1 A, B**). For common motifs MM and UM, we perform Epigram contrasting CMRs and CURs. In short, Epigram looks for enriched motifs that best differentiate the foreground from the background sequences. In both sets of the input sequences, Epigram iterates through all possible *k*-mers to calculate their occurrences, enrichment over genomic background and enrichment over shuffled input. These values are combined to determine the enrichment of *k*-mers. Position weight matrices (PWMs) are then generated by first picking a top *k*-mer and enriched *k*-mers similar to itself to construct a seed PWM, which is then extended by adding more enriched *k*-mers that are a few base pairs shifted from the original one. The motifs are then further ranked and filtered based on how well they differentiate the foreground from the background using LASSO (least absolute shrinkage and selection operator) logistic regression. The final set of motifs is then evaluated by random forest.

For tissue-specific VMM and VUM, we contrasted top 6000 most methylated and unmethylated regions in each methylome. In total, we identified 5172 motifs from 35 Epigram runs (34 methylome + 1 common) with default parameters[11] before curation (**Figure 2.1 C**). For each run, Epigram found DNA motifs that discriminate enrichment peaks of the high methylation region under consideration (e.g. CMR) from a background of low methylation region (e.g. CUR). Importantly, the background has the equal GC content, the number of regions and sequence lengths as the foreground to avoid inflated prediction results caused by simple features or unbalanced data set.

2.2.2 Motif curation and defining motif occurrence site

Following our previous study[11], we match motifs to the 1156 known motifs documented by the HOCOMOCO ChIP-seq consortium[104] using an E-value cutoff of 0.05 with Tomtom[105]. Next, we merged the similar motifs to remove redundancy. We calculated a pairwise motif distance using weighted Jensen-Shannon Divergence:

$$Distance = \sqrt{\frac{\sum_{k=0}^{nAli-1} JSD\left(M_1\left(i+k\right), M_2\left(j+k\right)\right)^3}{nAli}} + G\left(nAli, nGap\right)$$

$$G\left(nAli, nGap\right) = \frac{gapP * nGAP^2}{nAli}$$

where M_1, M_2 are PWMs of the two motifs, respectively, *M(i)* represents the ith column in the matrix, *JSD(x,y)* is Jensen-Shannon divergence, *nAli* and *nGAP* are respectively the lengths of the aligned sequence and gaps. Gap penalty function *G* has *gapP* as weight parameter set at 0.1. To ensure high similarity within the motif cluster, the gap penalty function is set to quadratic which is more stringent compared to traditional linear function to prevent having excessive gaps and hangovers. Motifs were hierarchically clustered with UPGMA[106] algorithm and clusters were chosen using a distance cutoff of 0.1. As a result, we obtained 3226 clusters and selected the motif closest to the centroid of the cluster to represent all the motifs in that cluster. We combine the *P*-value of motifs in the cluster using Fisher's combined probability test. Enrichment of each cluster is combined by geometric mean. Each unique motif is named by its group (MM or UM), combined *P*-value (log), combined enrichment, number of similar motifs in the cluster, followed by a short descriptive string. This string is either its best aligned known motif (e.g. UM_180.0_3.14_0.56_7_known-CTCF) matched by Tomtom described previously, or a consensus sequence (e.g. MM_10.2_2.16_0.54_1_ATKGCGSCA) determined by a minimal information loss method[10]. The strongest 313 motifs are filtered by volcano test with combined $P < 1e-10$ and *enrichment* > 2 (**Figure 2.6**). Finally, motif occurrence sites are determined by a $P < 1e-5$ calculated by FIMO[107].

2.2.3 Normalized motif occurrence and center-to-edge enrichment at DNMTs and TETs ChIP-seq peaks.

DNMTs and TETs occurrences were downloaded from the published studies[27, 108–110], including the ChIP-seq peaks of TET1 in HuES8 (a human embryonic stem cell line) from Verma et al.[110], TET2 in HEK293T (a human embryonic kidney cell line) from Suzuki et al.[27], TET2/TET3 in HEK293T from Deplus et al.[108], and DNMT1/3A/3B in NCCIT (a human embryonic carcinoma cell line) from Jin et al.[109]. The 5000bp neighbor regions around the ChIP-seq peaks were included as the background or edge. Normalized motif occurrence was calculated using the following formula.

$$NormalizedMotifOccurrence = \frac{Observed\left(MotifOccurrence\right)}{Expected\left(MotifOccurrence\right)}$$

$$Expected\left(MotifOccurrence\right) = \frac{TotalMotifOccurrenceLength * TotalChipSeqPeakNumber}{GenomeSize/BinWidth}$$

where $Observed\left(MotifOccurrence\right)$ is the observed occurrence number of a motif in a 100 bp bin, $TotalMotifOccurrenceLength$ is the total length of genome-wide motif occurrences defined by FIMO (see the above section), $TotalChipSeqPeakNumber$ is the total number of ChIP-seq peaks, $BinWidth$ is 100 bp and $GenomeSize$ is the genome size of 3.14E9 bp for the human genome hg19. We did this calculation for each of the 313 top enriched motifs in each 100 bp bin. We also downloaded 6251 differential CpGs (dCpGs) with $P < 0.05$ defined by Kemp et al[111], which were CpGs showing destabilized methylation level when CTCF contains point mutation or copy number aberrations. Center-to-edge enrichment of motif occurrences in the 500 bp around these reported dCpGs was performed the same as described above. Results are plotted in **Figure 2.2C and 2.7B**.

Further, center-to-edge enrichment was calculated by $NormalizedMotifOccurrence$ in the center 100 bp ChIP-seq bin divided by the average of $NormalizedMotifOccurrence$ at the bins 2500 bp upstream and downstream. Average enrichment and standard deviation were calculated across all MMs or UMs, followed by a two-tailed two-sample t-test, with $P < 0.01$ marked as significant. Results are plotted in **Figure 2.2D**.

2.2.4 Quantitative trait loci (QTL) enrichment analysis with TCGA

We downloaded the processed data (level 3) of 36 TCGA cancers from the Firebrowse service (http://firebrowse.org) including patient survival, somatic mutations, 450K methylation array, and RNA-seq data. All the somatic mutations taken from TCGA were first detected in Affymetrix Genome-Wide Human SNP Array 6.0, and determined by contrasting variants in cancer primary tissues with germline tissues, according to the TCGA Consortium[67]. Matrix eQTL[112] linear model was used to identify mQTL (methylation quantitative trait loci) and eQTL (expression quantitative trait loci) co-variating with methylation and transcript RNA-seq level, with 5000 bp distance cutoff from somatic mutation to CpG and transcript TSS, respectively. We used a conservative P-value cutoff of 0.01 on top of an FDR cutoff of 10% . Then we calculated the number of mQTL or eQTL out of all somatic mutations in 10 bins of gene body, i.e., 0-10% , 10-20% , . . . , 90-100% of the mRNA transcript length, defined in Gencode v19[113]. We performed such analysis on all genes and repeated it with the UM and MM occurrence sites (**Figure 2.2A**). To determine the significance of QTL enrichment, a chi-square test was carried out in each of the 10 bins of gene body, with the null hypothesis that mQTL% or eQTL% occurring at motif sites are the same as the rest of all genes, $P < 0.01$ are marked as significant.

2.2.5 Methylation quantitative trait loci (mQTL) enrichment analysis with three independent datasets

Three human methylome studies with independently called mQTLs were collected, i.e. human life course study[114], GenCord Cohort study[115] and a Schizophrenia study[116]. We took the mQTL SNPs identified from the original studies and these can be either somatic or germline mutations, which were not distinguished in the publication work. In total, there are around 16,000 to 30,000 identified mQTLs collected from these published studies. We defined an enrichment score using the following formula.

$$EnrichmentScore = \frac{Observed\,(mQTLOccurrence)}{Expected\,(mQTLOccurrence)}$$

$$Expected\,(mQTLOccurrence) = \frac{TotalmQTLOccurrence * TotalMotifOccurrenceLength}{GenomeSize}$$

where $Observed\,(mQTLOccurrence)$ is the observed occurrences of mQTLs in the occurrence sites of the 313 motifs genome-wide, $TotalmQTLOccurrence$ is the total number of mQTLs identified in each study, $TotalMotifOccurrenceLength$ is the total length of genome-wide motif occurrences, and $GenomeSize$ is the genome size of 3.14E9 bp for the human genome hg19. The occurrences of motifs have been defined by FIMO (see the above section).

We repeated this process in all samples from all three studies and calculated the standard deviation. Specifically, (1) 5 life stages from birth, childhood, adolescence, pregnancy and middle age in human life course study (blood samples from 1018 mother-child pairs), (2) 3 tissues from fibroblasts, LCLs and T-cells in GenCord cohort by Maria *el al* (204 newborn umbilical cord samples) and (3) 3 regions from prefrontal cortex, striatum and cerebellum of adult brain regions in the Schizophrenia study (173 fetal brain samples ranging from 56 to 169 days post-conception).

Finally, we used a single-tail one-sample t-test to determine the statistical significance ($P < 0.01$, **Figure 2.8A**).

2.2.6 Predicting TCGA cancer type with somatic mutation and motif

For each of the 32 TCGA cancers (in total 7120 patients), we trained two gradient boosting models[117] (mutation and mutation+motif) to distinguish one specific cancer from the other cancers. We chose gradient boosting implemented in Scikit-learn[118] and tuned its parameter based on a recent study[119], which showed that this decision tree based model is robust and performs well. Note that TCGA has 4 aggregated cancer types (GBMLGG, COADREAD, KIPAN and STES) that combine individual cancers such as GBMLGG combining GBM and LGG; we excluded them from the 32 TCGA datasets to avoid inflating the performance due to using the same patients in both the training and testing sets. In a mutation-only model, the cancer subtype of each patient was predicted only by somatic mutations as features. Because the input features are large (1.3 million unique somatic mutations for 7120 patients), we first reduced feature number. Each feature was assigned a score by the gradient boosting out-of-bag importance and averaged in 5-fold cross-validation to avoid overfitting. Features with negative importance scores were removed. The optimal number of features were determined as we observed the best model performance at around 500 features (**Figure 2.9A, upper panel**). Top 500 somatic mutations ranked by the average score were used while assuring equal or better performance compared to the full model (**Figure 2.9A, lower panel**).

After feature selection, we obtained 500 selected somatic mutations (from here referred simply as mutation). We used a series (length 500) of 0s and 1s to indicate which mutations a patient has. For example, 1,1,0,1, ... indicates patient have the 1st, 2nd and 4th mutation. For a mutation+motif model, each patient was represented not only by these 500 selected mutations but also by whether each of the 313 motifs is disrupted by mutations. We used a series (length 313) of integers to indicate how many mutations (without feature selection) are harbored in the occurrence sites for each of the 313 motifs. For example, 10, 20, 0, ... indicates there are 10 mutations in all the occurrence sites of the first motif, 20 in the second and none in the third. The performances of the two models were evaluated by auROC and auPRC with 5-fold cross-validations for each cancer (**Figure 2.4A**). Feature importance was determined by the default out-of-bag (OOB) important scores using the mean decrease of Friedman squared error over all cross-validated predictions in mutation+motif models. We filtered features with importance score > 0.01 within the enriched 313 motif groups and mutation located in the well-studied driver genes identified by the IntOGen Consortium[120]. To reduce false positives of selecting predictive features, we only considered 26 out of 32 TCGA cancers that showed $auPRC > 0.3$ (**Figure 2.4B**).

2.2.7 Predicting TCGA patient survival with somatic mutation and motif

All patients in 22 TCGA cancers with patient survival and mutation information were dichotomized based on 5-year survival to train two gradient boosting models (mutation and mutation+motif). We used the same 500 mutation features and 813 mutation+motif features from the diagnosis analysis and cross-validations were performed the same way as described above. The model performance was evaluated by the log2 hazard ratio and Kaplan-Meier estimator of the patient 5-year survival rate in the R package survival[121] (**Figure 2.4C**). Multivariate survival analysis was performed to show factors significantly ($P < 0.05$) correlated with patient survival with 95% confidence interval (**Figure 2.4D**).

2.2.8 Feedforward loop analysis

We built a network with three types of nodes: motifs, TET1/DNMT3A, coding genes. We have defined promoters as the region -1000bp and +500bp from the transcription start sites (TSS) of protein-coding genes (including TET1 and DNMT3A) from Gencode v19, as previously described. A directed edge is defined if the source node has an occurrence site at the promoter of the target nodes. For TET1 and DNMT3A, occurrence site is defined by ChIP-seq data previously measured in hESC and NCCIT cells, respectively. For motifs, the occurrence site is defined by FIMO with $p < 1e-5$. When a coding gene is a target, we first check if the gene is a known transcription factor, then define its binding site by FIMO with $p < 1e-5$. Finally, tracks are visualized in integrated genome viewer and the methylation track is provided by WGBS of H1 from The Epigenomics Roadmap Project.

2.3 RESULTS

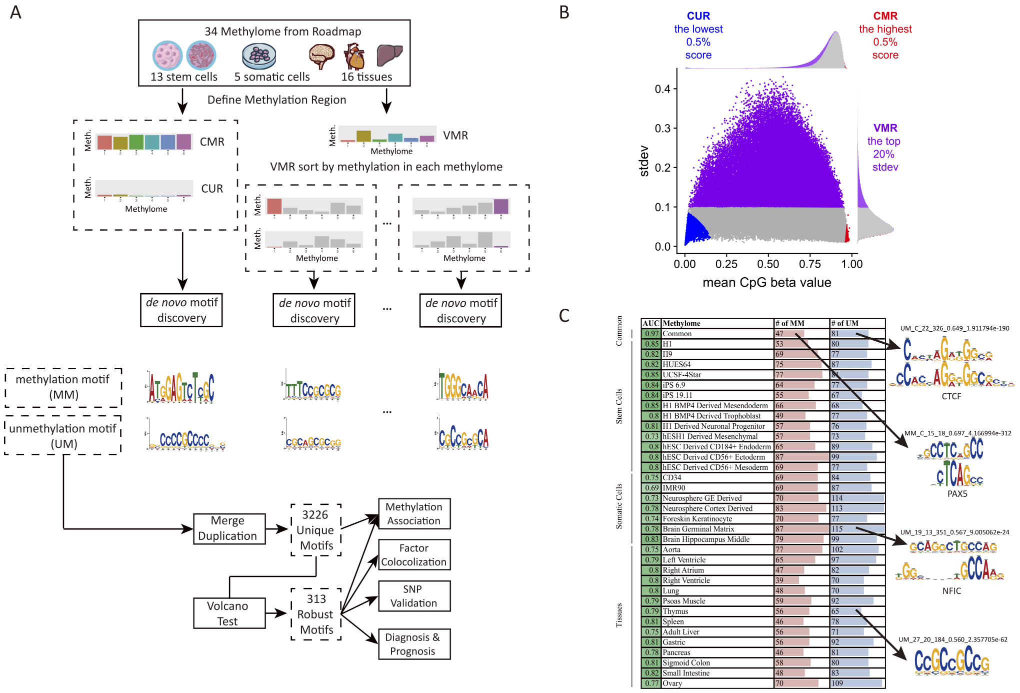
2.3.1 Defining DNA methylation regions and the de novo motif discovery

We aimed to identify DNA motifs associated with DNA methylation and thus started with searching for methylation regions that have the strongest signals. We collected whole genome bisulfite sequencing (WGBS) data of 34 human methylomes generated by the NIH Roadmap Epigenomics Project[122] (**Figure 2.1A**). We took an approach similar to the Ziller *et al.* study[2] and defined 1.55 million methylation regions containing 11.5 million CpG sites in the 34 methylomes. Because the methylome data is noisy, we only considered regions containing 2 or more CpGs within 400 bp apart, which covers 29.2% of the human genome.

Methylation level is associated with different functions. For example, low methylated regions (LMRs) are important in hematopoiesis and leukemia development[123], DNA methylation valleys (DMVs) are long hypomethylated regions involved in embryonic development and tissue-specific regulation[124, 125]; focal hypermethylation and long-range hypomethylation are found in cancer[126]; variably methylated regions (VMR) are associated with histone modification and enhancer[127] In this study, we defined three types of methylation regions based on the mean and standard deviation of the CpG methylation level in each region (**Figure 2.1A, B**): (1) Top 0.5% (or 7726) commonly methylated regions (CMR) which have the highest methylation level across 34 methylomes; (2) Top 0.5% (or 7726) commonly unmethylated regions (CUR) with the lowest methylation levels; (3) Top 20% (or 309040) variably methylated regions (VMR) with the highest standard deviation and this percentage is consistent with the previously reported 21.8% to 22.6% VMRs in the methylome[2, 127]. We are aware that these regions can vary upon the data sets used to define them. Because the 34 methylomes are derived from diverse cells and tissues, we argue the derived motifs are still reasonable starting points of revealing DNA binding proteins recruiting DNA methylation enzymes.

Defining commonly and variably methylated/unmethylated regions allow identification of motifs that are associated with DNA methylation independent of cell type or cell-type specific. CMRs and CURs are regions that show consistent methylation pattern across a diversity of 34 cells and tissues, and therefore they likely harbor motifs associated with methylation/demethylation in a cell-type independent manner. GREAT[128] analysis showed CMRs are strongly ($P < 1e-30$) linked to DNA repair and mitosis and are mostly (68%) found in introns (**Figure 2.6A**)[129]. CURs prefer promoters (66%) associated with ($P < 1e-30$) cell differentiation, development, and morphogenesis, indicating the important roles of demethylation in these processes[130, 131](**Figure 2.6A**). By contrasting CMRs to CURs, we identified 55 CMR and 87 CUR motifs using a motif finding algorithm Epigram[11] (**Figure 2.1A,C**). A 5-fold cross-validation using Epigram[11] successfully discriminated CMRs from CURs using the motifs ($AUC = 0.97$) (**Figure 2.1C**). Note that Epigram balances the GC content, sequence number, and length in the foreground and background, which avoids identification of trivial sequence motifs (see details in **Methods** and ref. [11]. Because these motifs are associated with high or low methylation regions commonly shared by diverse cell types, it is reasonable to argue that they are important or even casual for establishing, maintaining or removing DNA methylation.

Similar to TFs whose binding motifs are defined but their activities are specific, the usage of DNA methylation associated motifs is determined by the cellular state. The VMRs show cell type-specific methylation patterns, which provides an opportunity to identify motifs active in particular cell types. We contrasted top 6000 methylated and unmethylated VMRs sorted in each cell type and discovered average 63 methylation- and 85 unmethylation-associated motifs in each methylome, with an average AUC of 0.79 (**Figure 2.1C**).



AUC	Methylome	# of MM	# of UM
0.97	Common	47	81
0.85	H1	53	80
0.82	H9	69	77
0.82	HUES64	75	87
0.85	UCSF-4Star	77	81
0.84	iPS 6.9	64	77
0.84	iPS 19.11	55	67
0.85	H1 BMP4 Derived Mesendoderm	66	68
0.8	H1 BMP4 Derived Trophoblast	49	77
0.81	H1 Derived Neuronal Progenitor	57	76
0.73	hESH1 Derived Mesenchymal	57	73
0.8	hESC Derived CD184+ Endoderm	65	89
0.8	hESC Derived CD56+ Ectoderm	87	99
0.8	hESC Derived CD56+ Mesoderm	69	77
0.75	CD34	69	84
0.69	IMR90	69	87
0.73	Neurosphere GE Derived	70	114
0.78	Neurosphere Cortex Derived	83	113
0.74	Foreskin Keratinocyte	70	77
0.78	Brain Germinal Matrix	87	115
0.83	Brain Hippocampus Middle	79	99
0.75	Aorta	77	102
0.79	Left Ventricle	65	97
0.8	Right Atrium	47	82
0.8	Right Ventricle	39	70
0.8	Lung	48	70
0.79	Psoas Muscle	59	92
0.79	Thymus	56	65
0.81	Spleen	46	78
0.75	Adult Liver	56	71
0.81	Gastric	56	92
0.78	Pancreas	46	81
0.81	Sigmoid Colon	58	80
0.82	Small Intestine	48	83
0.77	Ovary	70	109

Figure 2.1: Defining methylated regions and searching for methylation associated motifs. **A**. The strategy of identifying DNA methylation associated motifs. **B**. WGBS CpG sites are merged within 400bp regions. Based on average CpG beta values of the region, we defined commonly methylated (CMR), un-methylated (CUR) and variably methylated regions (VMR). **C**. Identification of DNA methylation associated motifs in 34 cells and tissues. Example motifs are shown on the right (if matched to a known motif, the known motif logo is shown on the top).

In total, 5172 motifs were identified from 35 Epigram runs (1 common + 34 cell-specific). Because the same or similar motifs could be found in multiple cells, we clustered these motifs into 3226 unique ones using motif similarity measurement based on Jensen-Shannon divergence (see **Methods**). To control false discovery rate (FDR), we further conducted a robust volcano test[132] with a stringent requirement ($P < 1e-10$ and $enrichment > 2$), resulting in 313 methylation motifs for the follow-up analysis (**Figure 2.1A, 2.6B**), including 221 unmethylation motifs (UM) and 92 methylation motifs (MM). Among them, 36 (16.2%) and 14 (17.1%) are matched to 50 known motifs in the latest version of HOCOMOCO[104]. The matched included previously confirmed factors to influence methylation levels such as CTCF[23] and PAX5[133] as well as factors KLF4, SP4, and EGR1 that have been reported to regulate gene expression by binding to CpG rich promoters[134]. Furthermore, we also found 22 (24%) top enriched MMs were matched with the 657 reported methyl-specific motifs[70]. In addition, we have profiled the binding of 845 known TFs with ChIP-seq experiments documented in the latest GTRD (Gene Transcription Regulation Database)[135] in the motif occurrence sites (**Figure 2.7C**). These TFs can collaborate with the MMs/UMs to define the local methylation state. All motifs, their alignment results, and the TF occupancy profile can be found on our website (http://wanglab.ucsd.edu/star/MethylMotifs). The majority of the motifs are novel and showed strong sequence preference. UMs are more similar to each other and have higher GC content (e.g. CCGCCGCCG) than MMs (**Figure 2.6C,D**). Note that these motifs were found by Epigram after sequence balancing which removes GC content bias[11]. While high GC content and CpG-rich sequences have been associated with hypomethylation in regions such as CG-islands[63] and in specific cells[136–138], our analysis revealed specific DNA motifs with sophisticated patterns that may be recognized by proteins or ncRNAs.

2.3.2 Identified motifs are associated with the local DNA methylation deviated from the background

We first investigated the DNA methylation levels around the identified motif occurring sites (determined by FIMO[107] using $P < 10e-5$, the same parameters were used for all the relevant analyses thereinafter). We did observe hypomethylation and hypermethylation in the neighbor CpGs of the UM and MM motifs, respectively. Several representative examples are shown in **Figure 2.2A.** It is obvious that DNA methylation levels around the motif sites show a sharp ”dip” or ”peak”, suggesting the association is highly locus-specific. Interestingly, this trend remains the same in different cell types despite that the methylation levels in the surrounding regions vary. For example, motif UM_238.2_3.88_0.53_5 (matched to the WT1 motif) was identified from VMRs in the right ventricle tissues; the methylation level at its occurring sites decreases in all the cell types although the methylation level ranges from 0.6 to 0.8 in the surrounding regions (**Figure 2.2A**). This observation confirms the functionality of individual UM and MM motifs even though the local environment is overall hyper- or hypo-methylated.

We further examined the impact of these motifs on methylation in the gene coding regions. UM and MM consistently mark lower and higher local CpG methylation levels in the gene coding regions (**Figure 2.2B**). In the Roadmap dataset, we observed a significant impact of UMs or MMs on DNA methylation level around the transcription start sites (TSS) (**Figure 2.2B, left panel**). DNA methylation in the promoters is important for regulating gene expression[139] and thus itself is likely under active regulation. We observed the same trend in the TCGA DNA methylation data of 9037 patients from 32 cancers measured by Illumina 450K array[67] (**Figure 2.2B, right panel**). On average, CpG methylation decreases from the beta value of 0.81 in the Roadmap dataset, dominated by normal cell lines and tissues, to 0.59 in the TCGA cancer patients across 20,260 protein-coding genes. This observation is consistent with the global hypomethylation in cancer cells that have been reported in the literature[124, 130, 140]. However, the MM and UM occurring bins still showed respectively higher and lower methylation levels than the background. As an example, UM and MM occurrence sites are characterized by lower and higher methylation in the gene coding region of TP53 (chr17:7,540,000 - 7,650,000) in both TCGA and Roadmap data. Collectively, our results on two separate data sets generated by different technologies support that the identified DNA motifs play critical roles in influencing the local CpG methylation.

2.3.3 Identified motifs are significantly enriched at TETs and DNMTs binding sites

Locus-specific DNA demethylation or methylation depends on the recruitment of specific enzymes such as TET[9] and DNMTs[8] to particular genomic regions[5–7]. We reasoned that, if

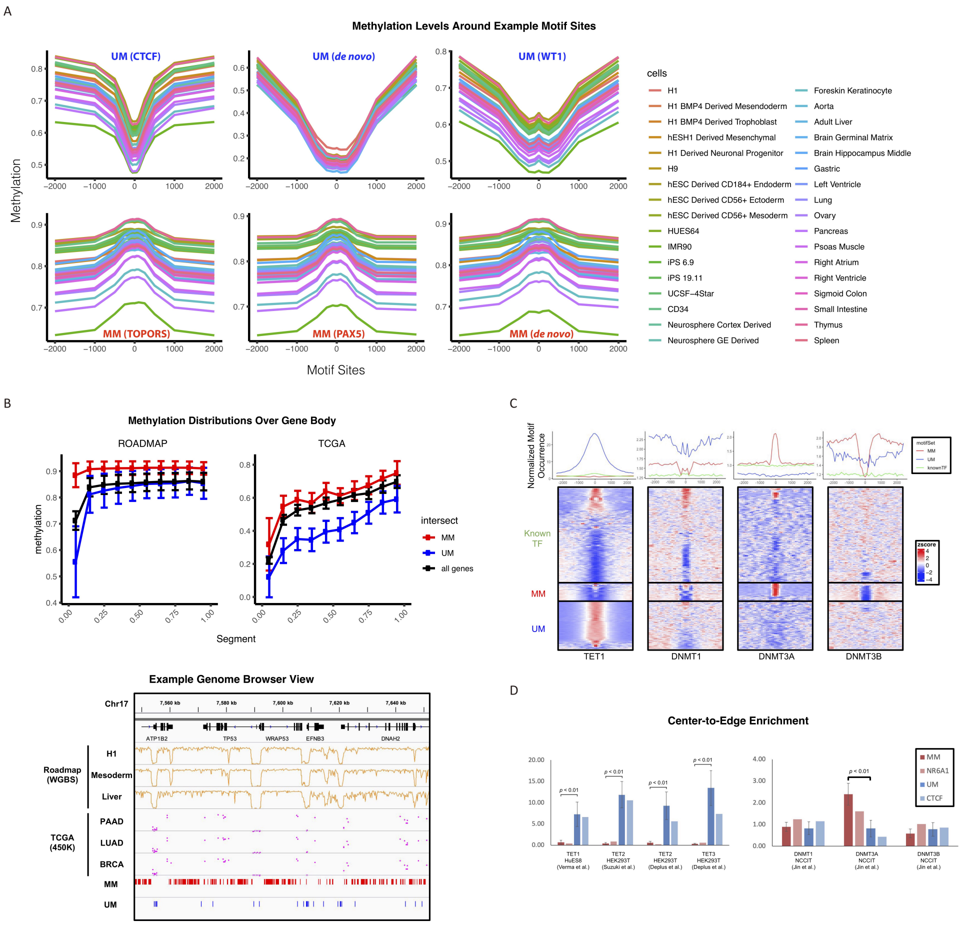


Figure 2.2: Identified motifs mark methylation level. **A.** Example motifs are shown with average CpG methylation level calculated in 50 bp bins around all motif sites, determined by FIMO at 1e-5 *P*-value cutoff. The examples are chosen to minimize bias and include a variety from both MM and UM, *de novo* and matched known TFs, common region and sorted variable regions. Upper panel, from left to right: UM_180.0_3.14 (matched to CTCF); UM_106.1_4.08 (*de novo*); UM_238.2_3.88 (matched to WT1); lower panel, from left to right: MM_65.9_2.90 (matched to TOPORS); MM_814.4_2.02 (matched to PAX5); MM_206.3_2.16 (*de novo*). **B**. DNA methylation levels in the ROADMAP (left) and TCGA (right) data sets over the gene body. Each gene body was split into ten equal bins and the Beta values of all CpGs in the same bin were averaged over all genes. Lower panel shows the correlation between the motif occurrences and CpG methylation in ROADMAP (WGBS data from H1, mesoderm, and liver) and TCGA (450K methylation of CpGs averaged in patients from PAAD, LUAD, and BRCA) around TP53 (chr17:7,540,000 - 7,650,000). **C.** Normalized motif occurrence of UM, MM and known TFs (excluding matched) from HOCOMOCO[104] at 5000 bp windows centering ChIP-seq peaks of TET1, DNMT3A and DNMT3B collected from various studies[27, 109, 110]. The lower panel shows the clustered heatmap of normalized z-score. **D.** Center-to-edge enrichment of UMs and MMs in comparison with TF NR6A1 and CTCF, which were reported to recruit DNMT and TET to specific loci, at the ChIP-seq peaks of DNMTs and TETs.

the identified motifs are important for recruiting the enzymes, these motifs would be enriched around the binding sites of the recruited enzymes. To this end, we have collected all the available ChIP-seq experiments of TET and DNMT enzymes[27, 108–110]. Indeed, at the center of TET1 ChIP-seq peaks in hESC H1 cells[110], the UM sites occur 26.7 times of expected counts (see details in **Methods**), whereas MM motifs occur roughly same (1.4 times) as the expected counts (**Figure 2.2C**, the first panel from the left). This observation is consistent with the previous reports that TETs can be recruited to specific locus by DNA binding factors[7, 9]. Interestingly, the wide distribution of UM around TET peaks compared to MM-DNMT overlap is consistent with the previously reported role of TET in protecting spanned low-methylation regions termed methylation canyons against hypermethylation[141]. Furthermore, TET prefers CpG-rich patterns such as CpG island which spans several kilobases[24] and can bind CpG-rich DNA sequences[5] in mammalians to maintain stable demethylation[25]; consistently, UMs have significantly higher GC content than MMs and known motifs ($P < 0.05$, **Figure 2.6C**).

We observed different motif occurring patterns around the binding sites of different DNMT enzymes. DNMT3A and DNMT3B are responsible for *de novo* methylation[142]. At the center of DNMT3A ChIP-seq peaks in the human NCCIT cells[109], we observed a peak of the MM motif occurrence compared to the known and UM motifs (**Figure 2.2C**). Interestingly, the MMs are enriched at the shoulder regions of the DNMT3B binding sites but depleted at the center (**Figure 2.2C**). Note that only 2.2% of DNMT3A and 3.8% of DNMT3B peaks overlap with each other[109] (**Figure 2.7A**). Several studies have demonstrated some distinct roles of DNMT3A and DNMT3B, showing that DNMT3B preferentially targets gene bodies marked with H3K36me3[57, 143–145]; in fact, H3K36me3 is 4.27 times enriched at the DNMT3B compared to DNMT3A peaks in gene coding regions (**Figure 2.7A**). These observations suggest that the MMs are likely recognized by DNA binding factors involved in actively recruiting DNMT3A, whereas DNMT3B may be recruited by flanking sequences containing MMs and together with chromatin marks and/or other factors such as H3K36me3. Interestingly, DNMT1, an enzyme involved in DNA methylation maintenance and recognizing hemimethylation[37], shows a different profile from DNMT3A/B (**Figure 2.2C**, second panel from the left). This difference may have resulted from the different mechanisms or factors involved in active and passive DNA methylation.

To further validate if the observed co-occurrence around methylation enzyme is significant, we also compared the center-to-edge enrichment of UM and MM with TFs known to regulate DNA methylation (**Figure 2.2D, method**). Previous studies have reported that introducing a CTCF binding site at a particular locus leads to local DNA demethylation and enrichment of TET[23]. NR6A1 has also been confirmed to recruit DNMT to methylate at target genes[37]. Here, we show that at the center of TETs binding sites, UMs are significantly more enriched than MMs, and have even higher enrichment than CTCF (**Figure 2.2D**, left panel). Similarly, MMs are significantly more enriched than UMs at the center of DNMT3A binding sites, surpassing that of NR6A1 (**Figure 2.2D**, right panel). The enrichment of MMs and UMs were further compared with the known TFs such as PAX5, TOPORS, WT1 and PPARG that are most enriched at the TETs and DNMT3A sites. Furthermore, we downloaded the most confident ($P < 0.05$) differential CpGs (dCpGs) defined by Kemp et al[111], i.e. CpGs showing destabilized methylation level when CTCF contains point mutation or copy number aberrations in several human cancers. The CTCF's critical role in affecting the local DNA methylation in these loci was confirmed and we indeed found that CTCF and UMs were even more enriched at these loci (**Figure 2.7B**). These results demonstrated that the identified motifs can be recognized by particular DNA binding factors that in turn recruit the methylation modifying enzymes in a locus-specific manner. Given that the majority of MMs (71.4%) and UMs (83.9%) are *de novo* motifs, our findings pave the way towards identifying particular factors involved in locus-specific methylation regulation.

2.3.4 Genetic variants at identified DNA motif sites are associated with altered methylation level

To validate the functionality of the identified motifs, we investigated the enrichment of quantitative trait loci of expression (eQTL) and methylation (mQTL) at motif occurrence sites. Note that we only took somatic mutations identified by the TCGA consortium in this analysis. We analyzed the relationship between somatic mutation and methylation level using the TCGA data[67] and identified methylation quantitative trait loci (mQTL), which are somatic mutations correlating with CpG variation within 5000bp. Using Matrix eQTL[112], we identified 26341 mutation-CpG pairs (mQTL), corresponding to 17038 unique mutations and 20043 CpGs, from a total of 1.3 million somatic mutations in 9037 patients of 32 cancers. We observed an average 11.7% mQTL discovery rate at the motif sites compared to 2.3% in the background (**Figure 2.3A, upper left panel**). This enrichment difference is most prominent around the transcription start site, suggesting that the identified motifs have a stronger impact on methylation at TSS (**Figure 2.2B**)[96, 146, 147]. Enrichment of mQTL in both MM and UM sites was also found in three additional human methylome datasets using the reported mQTLs in the original studies[114–116] (**Figure 2.8A**), which confirms the generality of this observation. Because DNA methylation is associated with gene expression[1, 2], it is not surprising that MMs and UMs significantly overlap with expression quantitative trait loci (eQTL), which are mutations correlated with gene expression level (**Figure 2.3A, right panel**).

To investigate the causality between these motifs and DNA methylation level, we analyzed whether disrupting these motifs would lead to DNA methylation change. We chose to focus on the possible binding sites of TET1 and DNMT3A containing these motifs because the significant enrichment of the found motifs in the enzyme-binding regions implies that the active methylation/demethylation is most likely mediated by DNA binding factors to recruit TET1/DNMT3A. Despite the ChIP-seq experiment of TET1/DNMT3A was done in one particular cell type, the

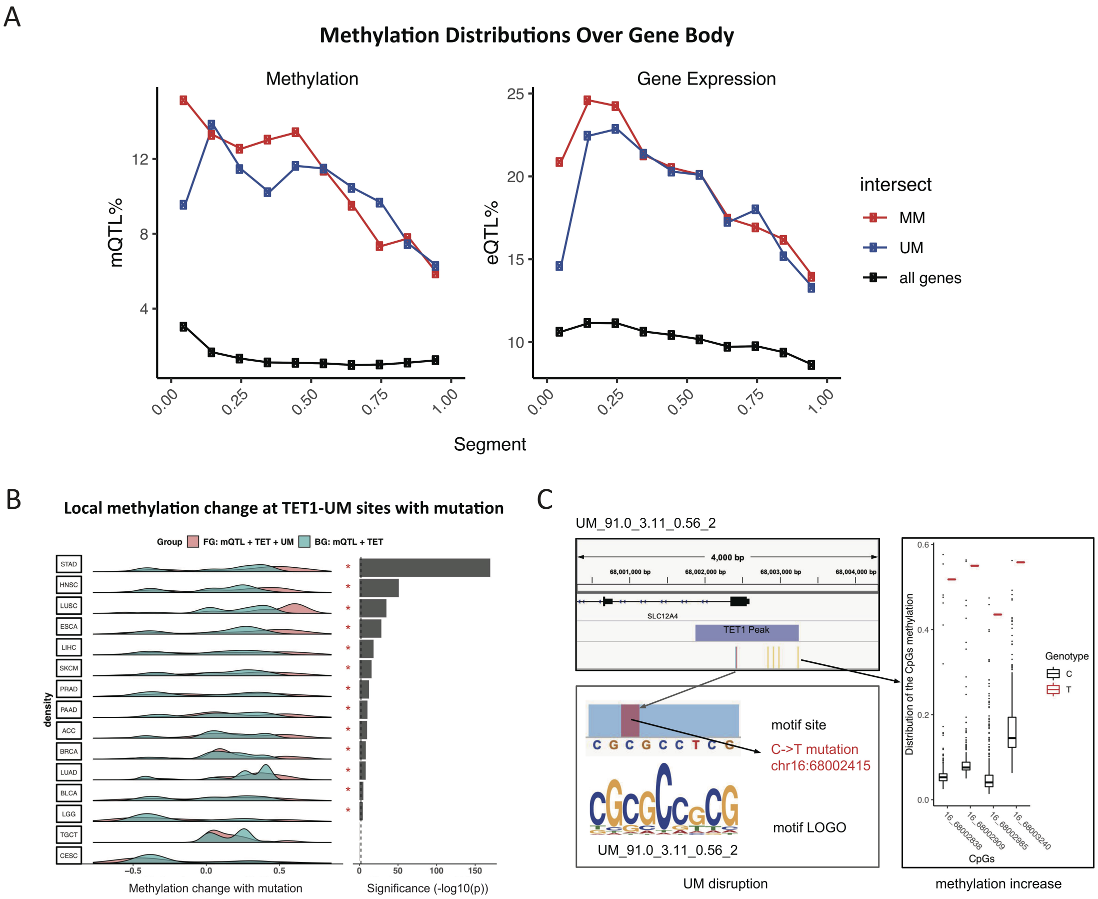


Figure 2.3: Somatic mutation at motif sites co-occur with local methylation alteration. **A.** Distribution of quantitative trait loci corresponding to methylation (mQTL) and gene expression (eQTL) over gene body (see details in **Methods**). Each gene body is split into ten equal bins. **B.** Methylation level change of CpG sites nearby TET1-UM sites (TET1 binding peaks containing UM motifs) overlapping with somatic mutations. Asterisks indicate $P < 0.01$ calculated with paired one-tail t-test, pairing foreground observed methylation change to the corresponding background expected methylation change. Foreground (FG), mQTL at TET1-UM sites. Background (BG), mQTL at TET1 binding peaks[27, 109, 110]. To ensure the statistical significance, we only considered the 15 cancers with > 100 CpGs within 5000bp of TET1-UM sites (see details in **Methods**). **C.** An example showing disruption of a UM motif (no match with known motifs) by a C>T somatic mutation at chr16:68002415 significantly increases the methylation level of the 4 nearby CpGs in the LUAD patients.

sequence features, i.e. the motif composition in these regions, do not change and thus the mechanism of the active methylation regulation. The methylation change is decided by which factors are expressed and active in a specific cell type or state. Disrupting these motifs would lead to methylation change in the nearby CpGs.

Using the TCGA data, we first identified 5372 CpG sites from 15 cancers within 5000bp of the TET1 binding peaks that also contain mutations overlapping with UMs in at least one patient. Because we did not have TET1 ChIP-seq data in the cancer patients, we used the published data measured in hESC (see **Figure 2.2C, D**). We compared the methylation change of these CpGs between patients with and without the mutation in each cancer. 13 out of 15 cancers showed significant ($P < 0.01$) increased methylation level of with-mutation compared to the without-mutation patients (background) (**Figure 2.3B**, see **Methods** for details). One example is given in **Figure 2.3C** for a UM motif UM_91.0_3.11_0.56_2. This motif is within a TET1 peak and is disrupted by a C>T somatic mutation at chr16:68002415 on the first exon of SLC12A4 in one LUAD cancer patient. All 4 CpGs within 500 bp upstream of the mutation showed increased methylation (beta value increased from 6.2% to 52% , 8.8% to 55% , 6.2% to 44% and 17% to 56% , respectively). Hypomethylation in the SLC12A4 promoter is related to resistance to platinum-based chemotherapy in ovarian cancer[148]; the 4 CpGs affected by the mutation are located in the SLC12A4 promoter, suggesting a mechanism of how the mutation may affect response to chemotherapy through regulation of local DNA methylation. More examples of mutation-induced methylation change through disrupting UMs are shown in **Figure 2.8B.**

Overlapping MM and mutations with DNMT3A peaks only resulted in $<$ 100 CpGs sites in 2 cancers. Although we observed decreased methylation level of DNMT3A-MM overlapping with mQTL as predicted, the analysis did not have enough statistical power. Because the methylation was measured by 450K array and mutations were detected and called from Affymetrix Genome-Wide Human SNP Array 6.0, it is reasonable to expect that more sites can be observed with whole methylome and whole genome sequencing data.

2.3.5 Combining Motifs and somatic mutation Shows Diagnosis and Prognosis Power

DNA methylation has been shown to be predictive for cancer diagnosis and patient survival prospective[76, 93]. Since we have shown motif disruption is associated with methylation change, we hypothesized that combining motifs with mutations can improve prediction for cancer diagnosis and patient survival. To evaluate this, we trained gradient boosting models[117] using mutation and mutation+motif as features in 32 TCGA cancers from 7120 patients (see **Methods** for details). We calculated both auROC and auPRC (a metric for an imbalanced dataset to avoid inflated evaluation of the performance)[149]. The inclusion of the motifs in the models resulted in increased auROC and auPRC in all the 32 cancers. On average, auROC increased from 0.78 to 0.92 and auPRC from 0.45 to 0.56, whereas 26 (for auROC) and 13 (for auPRC) improvement are statistically significant ($P < 0.01$) (**Figure 2.4A**). Notably, several cancers showed drastic improvement, including ovarian cancer (OV, auPRC from 0.41 to 0.79), thyroid carcinoma (THCA, auPRC from 0.49 to 0.82), acute myeloid leukemia (LAML, auPRC from 0.6 to 0.88), pheochromocytoma and paraganglioma (PCPG, auPRC 0.49 to 0.75) (**Figure 2.9B**). These cancers all have reported aberrant methylome and have methylation associated diagnosis and therapeutic targets[150–153].

For 26 cancers with $auPRC > 0.3$, the 67 most predictive features ($score > 0.01$) determined by the gradient boost estimator are shown in **Figure 2.4B** (see **Methods** for details), including 13 mutations, 20 MMs, and 34 UMs. Only 2 MMs are matched to known motifs (RXRB and PAX5), whereas 7 UMs to AP2B, BTD, PLAL1, GLIS2, WT1, CNOT3, and GTF3A. The predictive mutations include those occurring on the cancer driver genes such as BRAF (in 16 cancers), TP53 (in 14 cancers), IDH1 (in 14 cancers), PIK3CA (in 13 cancers) and KRAS (in 12 cancers). Strikingly, we found numerous MMs and UMs very predictive in multiple cancers. Notably, MM_814.4_2.02_0.62_8 (PAX5) that has been shown to strongly impact local

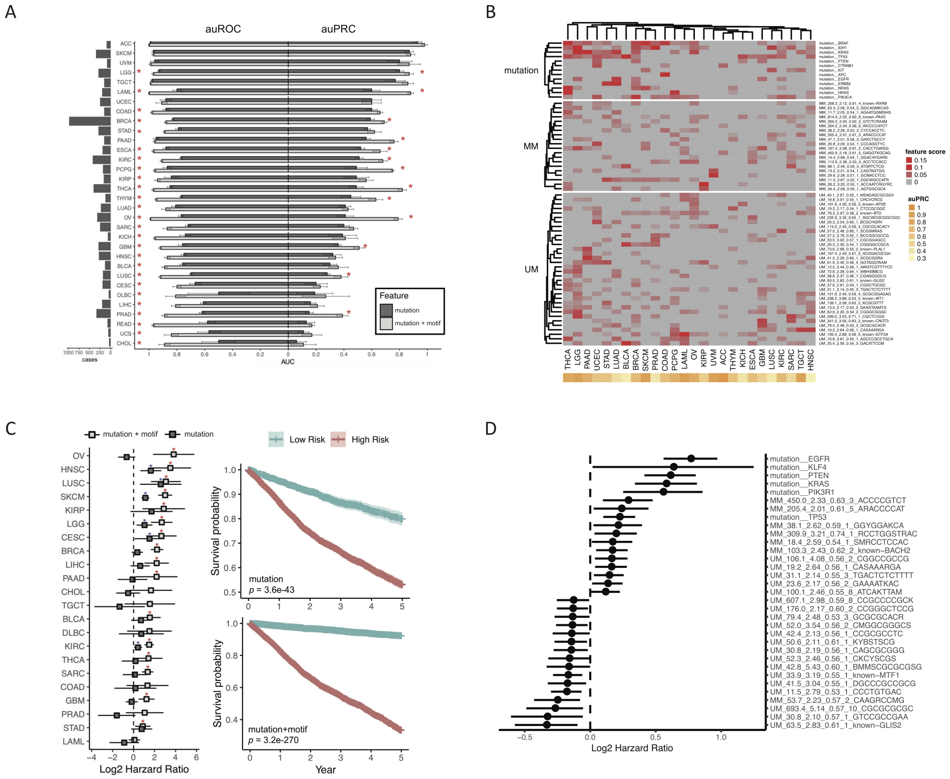


Figure 2.4: Combining motif to mutation improves the prediction of cancer subtype and patient survival. **A**. auROC and auPRC for cancer subtype prediction. Classification model of each cancer built with gradient boosting. Performance evaluated with auROC (area under the receiver operating characteristic, good for an overall evaluation.) and auPRC (area under the precision-recall curve, good for an unbalanced dataset where the positive label is scarce). Label: mutation: using somatic mutations as features. mutation+motif: using both somatic mutations and collective disruption of motif site as features (See **Methods** for details). $*$ Adjusted $P < 0.05$. **B**. Results of top predictive features ($score > 0.01$) using gradient boosting out-of-bag estimation. 26 cancers with $auPRC > 0.3$ are shown. **C**. Survival analysis with gradient boosting with mutation and mutation+motif as models. Left: multivariate survival analyses for all solid TCGA cancers. Forest plots showing log2 hazard ratio (95% confidence interval) of the predicted high-risk group by both models. $*$ Adjusted $P < 0.05$ (blue for the mutation model and red for the mutation+motif model). Right: Kaplan-Meier survival estimation (95% confidence interval) in the high-risk group versus low-risk group predicted by both models. **D**. Multivariate survival analysis showing factors correlating with patient survival ($P < 0.05$) with the log2 hazard ratio (95% confidence interval).

methylation level (**Figure 2.3C**) is important in 12 cancers. The 5 UMs predictive in > 10 cancers are UM_78.3_2.97_0.58_2 (BTD), UM_13.5_2.17_0.53_2, UM_195.4_2.88_0.56_5 (GTF3A), UM_35.4_2.56_0.54_3 and UM_61.9_2.40_0.56_4 (**Figure 2.4B**).

To evaluate the prognosis power of the motifs, we trained two gradient boosting models (mutation and mutation+motif) to discriminate low-risk from high-risk patients. We evaluated the performance using the survival hazard ratio of the predicted high-risk group (higher ratio means better performance). The mutation-only model found 6 out of 22 cancers having significant ($P < 0.05$) hazard ratio. In comparison, the mutation+motif model achieved 16 out of 22 cancers having significant ($P < 0.05$) hazard ratio (**Figure 2.4C, the left panel**, see **Methods** for details). Kaplan-Meier test showed a better separation of patient survival between the predicted low-risk and high-risk groups by considering motifs ($P = 3.6E - 43$ for the mutation-only model and $P = 3.2e - 270$ for the mutation+motif model, **Figure 2.4D, right panel**). Multivariate survival analysis on the full model revealed important factors correlated with patient survival ($P < 0.05$), including 6 mutations, 6 MMs and 20 UMs (**Figure 2.4D**). These results further confirmed the functionality of the discovered motifs and highlighted the potential for clinical application.

2.3.6 Motifs involved in both DNA methylation and histone modifications

Both DNA methylation and histone modification play important roles in regulating gene expression and their interplay has been well recognized[3, 4]. In a separate study, we identified 361 histone motifs[71] that are associated with 6 (H3K4me1, H3K4me3, H3K27ac, H3K27me3, H3K9me3, H3K36me3) histone modifications from 110 diverse human cell types/tissues. By comparing the 313 methylation motifs with these 361 histone motifs, we found that 56.5% MMs (52 out of 92) overlap with them (e-value cutoff of 0.05 using Tomtom) (**Figure 2.5A**). Among these, 35 MMs are aligned to H3K36me3 motifs as H3K36me3 can recruit DNMT3A/3B through their PWWP domain[154, 155]. In contrast, 74.2% (164 out of 221) UMs found no match to histone motifs. 57 UMs are matched to motifs associated with the active promoter or enhancer marks: 12 UMs matched to H3K27ac, an active promoter and enhancer mark; another 12 UMs matched to the promoter mark H3K4me3. As active enhancers and promoters tend to have low methylation[66], this observation is not unexpected. Interestingly, we observed another 12 UMs matched to the motifs associated with the poised promoter markers H3K4me3+H3K27me3. Previous studies also suggested the colocalization of H3K4me3 and H3K27me3 marks DNA hypomethylation in pre-implantation embryos[156].

2.3.7 Regulatory loops on DNA methylation

DNA methylation is dynamically regulated in response to the cell state change. We analyzed the putative regulatory connectivity between the identified motifs, transcription factors and the modifying enzymes of TET1 and DNMT3A. We only considered TET1 and DNMT3A here because their binding peaks are significantly enriched with UMs and MMs, respectively (**Figure 2.2C**). It is well accepted that a known TF motif occurring in the promoter of a gene suggests a possible regulation of the gene expression by the TF. Similarly, we infer the occurrence of a UM or MM in a gene's promoter indicates putative regulation on the DNA methylation level and thus affecting gene expression.

We first analyzed the promoters of TET1 and DNMT3A. We found 19 UMs in the promoters of both TET1 and DNMT3A. We also found these UMs appearing in the promoters of 25 TFs that also have motifs in the promoters of both TET1 and DNMT3A and presumably regulate the two enzymes (**Figure 2.5B**). Such a topology forms a feed-forward loop (FFL)[157] that involves three nodes: two regulator nodes (motifs and TFs), one regulates the other (motifs regulates TFs), and both jointly regulating a target (TET1 or DNMT3A) (see **Methods**). UMs induce demethylation of TET1/DNMT3A and their regulator TFs, which forms positive FFLs to enhance the expression of both TET1 and DNMT3A once the motifs are activated. We also found 2 and 5 MMs occurring in the promoters of TET1 and DNMT3A, respectively. These MMs appear in the promoters of 14 TFs as the other regulator of TET1 or DNMT3A, of which

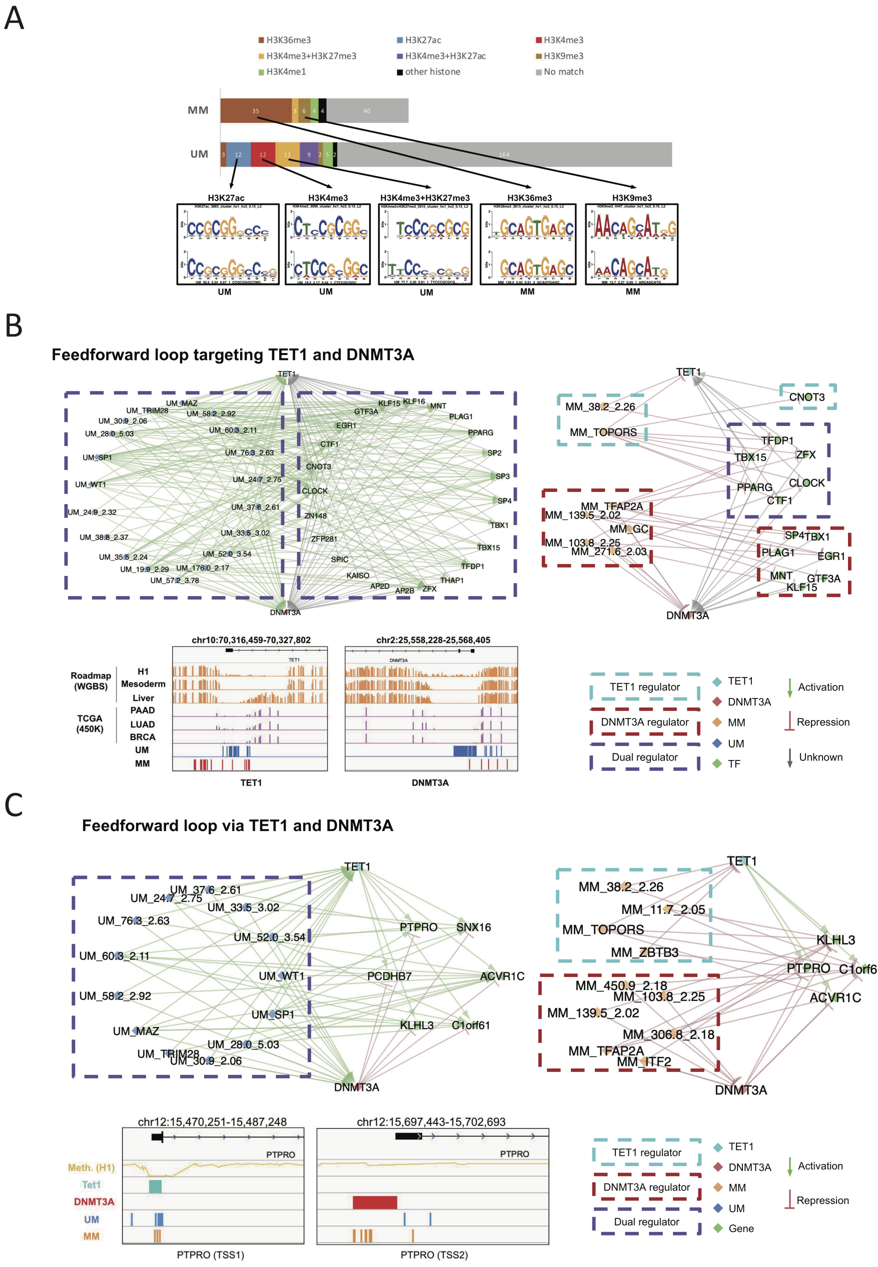


Figure 2.5: Methylation motifs interplay with TET1, DNMT3A, gene regulation, and histone modification **A.** Methylation motifs matched to histone motifs[71]. Motifs are aligned with Tomtom with $e < 0.05$. Lower panel showing several examples. **B.** Feedforward loop targeting TET1 and DNMT3A. **C.** Feedforward loop via TET1 and DNMT3A.

1 TF only regulates TET1, 7 TFs only regulates DNMT3A and 6 TFs regulate both (**Figure 2.5B**); these FFLs form enhanced dynamic regulation to repress TET1 and DNMT3A expressions. Overall, there are many more activating than repressive FFLs on regulating TET1 and DNMT3A.

Previous reports have also shown TET1 and DNMT3A have competitive binding to regulate promoters in mouse embryonic stem cells[158]. In addition, in honey bees, Dnmt and Tet (homolog of vertebrate DNMTs and TETs) were found to target memory-associated genes sequentially, while Dnmt3 was found in a negative feedback loop for DNA methylation[159]. We found 6 genes targeted by UMs and also by both TET1 and DNMT3A (as indicated by their ChIP-seq peaks in hESC and NCCIT cells, respectively) (**Figure 2.5C**). Interestingly, 4 of them (KLHL3, C1orf61, ACVR1C, PTPRO) are also targeted by MMs and either TET1 or DNMT3A (**Figure 2.5C**). One of them, PTPRO, a cancer suppressor and therapeutic target of a variety of solid and liquid tumors, is silenced by promoter hypermethylation[160]. In fact, we observed higher methylation at the promoter of the first TSS of PTPRO (TSS1, chr12:15,474,979-15,476,332) in the TCGA patients (beta value average at 0.15) compared to the Roadmap methylomes (beta value averaged at 0.05) (**Figure 2.5C**). PTPRO has multiple TSSs and alternative splicing forms[161], and each TSS has a TET1 or DNMT3A ChIP-seq peak (**Figure 2.5C**). As competitive binding of activator and repressor can lead to sharp turn on/off of the gene expression[162–164], we speculate the competitive FFLs formed by the motifs and modifying enzymes would thus allow dynamic regulation of the methylation and presumably the expression levels of these genes.

2.4 DISCUSSION

In this study, we present a comprehensive catalog of the DNA motifs associated with DNA methylation. We did observe coincident higher and lower methylation levels around the MM and UM occurring sites, respectively. Furthermore, the motif sites are also enriched with functional mutations, including mQTL and eQTL. We also showed that combining DNA motifs and mutations can achieve accurate prediction of diagnosis and prognosis in TCGA cancer patients, which supports the importance of these motifs.

Our analysis suggested that these motifs are most likely involved in recruiting TET and DNMT3A for active demethylation and methylation, as indicated by their significant enrichment in the binding sites of these enzymes. The passive or maintenance methylation mediated by DNMT1 seems to be regulated by mechanisms other than DNA binding co-factors because we did not observe an enrichment of the found motifs in the DNMT1 binding sites.

Interestingly, some of these motifs may also play roles in histone modifications as they were also found associated with histone modifications, particularly those relevant to DNA methylation such as H3K36me3 that were reported to recruit DNMT3A/B through their PWWP domains. Furthermore, these motifs can form feed-forward loops (FFLs) with TFs to regulate TET1 and DNMT3A or regulate genes together with TET1/DNMT3A. These FFLs allow possible regulation of the DNA methylation dynamics and presumably the gene expression dynamics. The interplay between DNA and epigenetic signatures is central to TF recruitment and eukaryotic gene expression regulation. Binding sites of TFs are determined by combined factors including DNA sequence, methylation[165], histone modification[166], and nucleosome landscape[167]. Our motif analysis suggests putative mechanisms for experimental test.

We have shown multiple lines of evidence to support that the identified motifs are involved in regulating DNA methylation. To confirm the causal relationship between TF-DNA binding and methylation, additional experimental tests are needed such as mutating the found motifs in a specific locus and measuring its impact on the local DNA methylation change. We have made all the motifs and their occurrence sites available, which will allow designing particular experiments for testing the functions of these motifs in disease or other biological contexts. These experiments are still challenging nowadays because it requires to simultaneously mutate multiple short motifs. Given the fast advancement of the genome editing technology, it will become feasible to perform such a test in a high-throughput fashion of the predicted motifs in the future. There exist more than one mechanism of establishing and maintaining locus-specific DNA methylation patterns[5, 165], which may require different combinatorial interactions between different factors. Our study establishes a catalog of the possible participating motifs, which provides a starting point towards fully deciphering the grammar of regulating the locus-specific DNA methylation.

Chapter 2, in full, is a reprint of the material as it appears in ”Identification of DNA motifs that regulate DNA methylation. Mengchi Wang, Kai Zhang, Vu Ngo, Chengyu Liu, Shicai Fan, John W Whitaker, Yue Chen, Rizi Ai, Zhao Chen, Jun Wang, Lina Zheng, Wei Wang.” in Nucleic Acid Research, 2019. The dissertation author was a primary investigator and author of this paper.

2.5 SUPPLEMENTARY

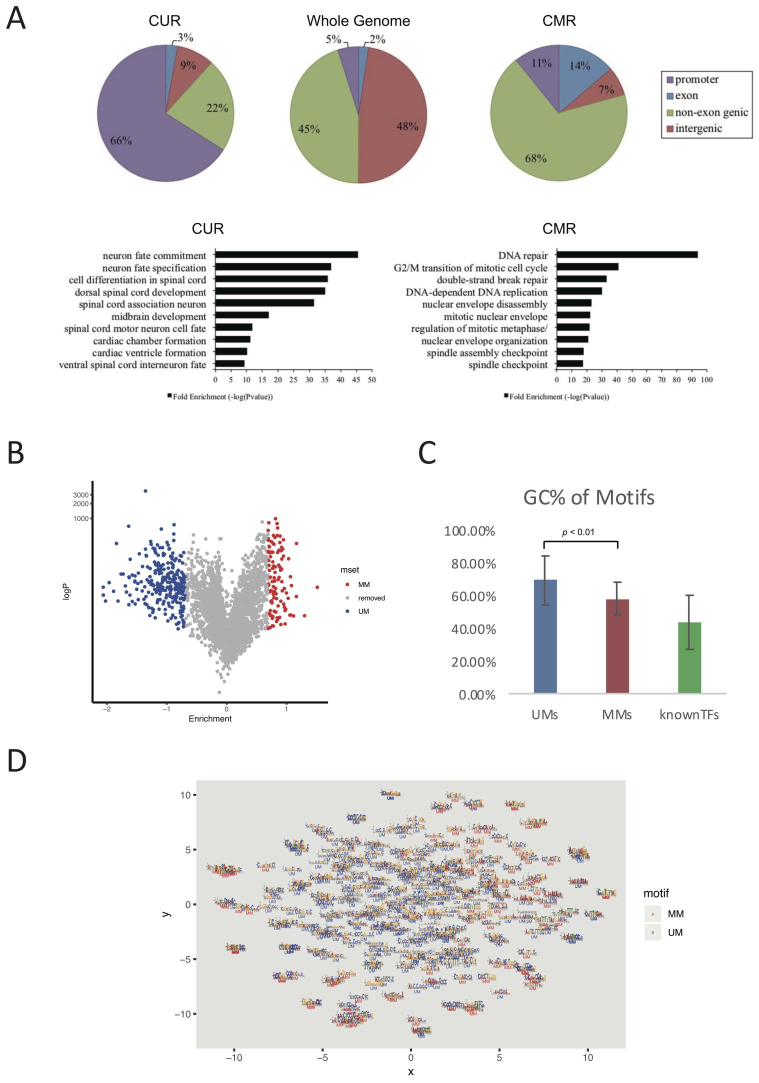
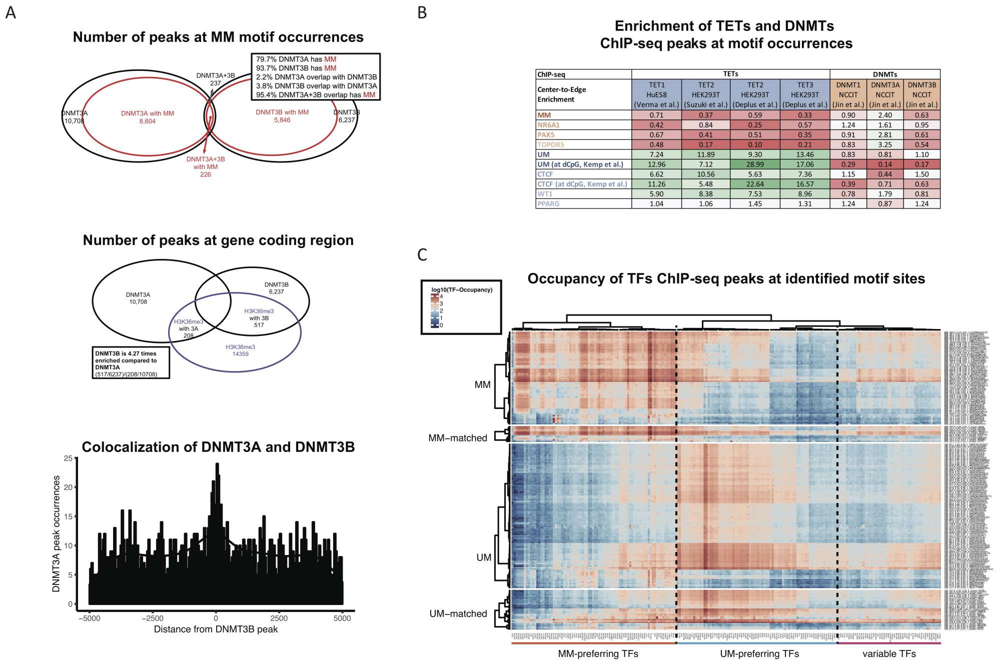


Figure 2.6: Characterization of the identified motifs and regions. **A.** Gene ontology analysis and genomic location of the CUR and CMR compared against the whole genome. **B.** Volcano plot of 313 top cluster filtered by $p < 1e - 10$ and fold-change $enrichment > 2$. **C.** The motif GC content of top 313 unmethylation motifs (UM), methylation motifs (MM) and known motifs curated from Hocomoco. **D.** tSNE plot showing sequence similarity among 313 motifs, pairwise distance calculation described in **Methods.**



ChIP-seq	TETs				DNMTs		
Center-to-Edge Enrichment	TET1 HuES8 (Verma et al.)	TET2 HEK293T (Suzuki et al.)	TET2 HEK293T (Deplus et al.)	TET3 HEK293T (Deplus et al.)	DNMT1 NCCIT (Jin et al.)	DNMT3A NCCIT (Jin et al.)	DNMT3B NCCIT (Jin et al.)
MM	0.71	0.37	0.59	0.33	0.90	2.40	0.63
NR6A1	0.42	0.84	0.25	0.57	1.24	1.61	0.95
PAX5	0.67	0.41	0.51	0.35	0.91	2.81	0.61
TOPORS	0.48	0.17	0.10	0.21	0.83	3.25	0.54
UM	7.24	11.89	9.30	13.46	0.83	0.81	1.10
UM (at dCpG, Kemp et al.)	12.96	7.12	28.99	17.06	0.29	0.14	0.17
CTCF	6.62	10.56	5.63	7.36	1.15	0.44	1.50
CTCF (at dCpG, Kemp et al.)	11.26	5.48	22.64	16.57	0.39	0.71	0.63
WT1	5.90	8.38	7.53	8.96	0.78	1.79	0.81
PPARG	1.04	1.06	1.45	1.31	1.24	0.87	1.24

Figure 2.7: Identified motifs interact with TETs, DNMTs, and TFs. **A.** methylation motifs co-occur with DNMT1 and DNMT3B ChIP-seq peaks in differentiated NCCIT cell, while unmethylation motifs co-occur with TET ChIP-seq peaks in human embryonic stem cells. The lower panel shows the histogram of DNMT3A peak counts 5000bp nearby DNMT3B peaks, with 50bp bin width. **B**. Details on center-to-edge enrichment of motifs and known TFs in respect to TETs and DNMTs ChIP-seq peaks. Differential CpGs (dCpGss) are identified in Kemp et al, 2014[111] (see **Methods.**). **C.** Occupancy of TFs ChIP-seq peaks at identified motif sites. ChIP-seq peaks are acquired from the GTRD database. Motifs occurrence sites are defined by FIMO (see **Methods.**). Occupancy is defined by the total number of peaks at the 500bp windows centered at occurrence sites of each motif. Clustering is performed by using Euclidean distance and the Ward method.

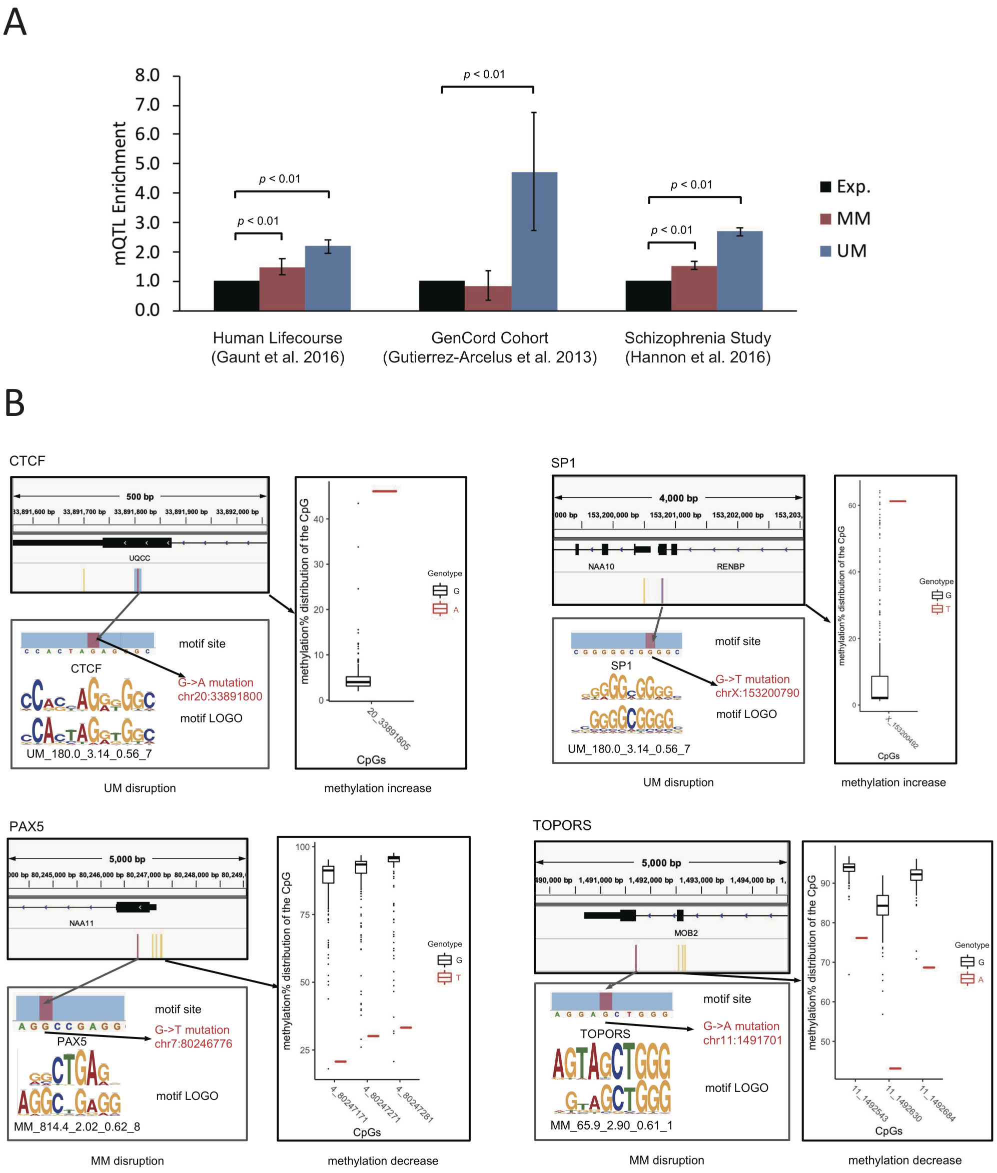


Figure 2.8: Identified motif occurrences overlap with TCGA functional SNPs. **A.** Enrichment of mQTL at UM and MM occurrence site using three additional studies, namely human life course study, GenCord Cohort study and a Schizophrenia study. Enrichment of mQTL are defined as observed mQTL ratio over expected ratio, with error bars showing standard deviation across samples (Left: across five stages in human life course; Middle: across three cell types; Right: across three adult human brain regions) and $p < 0.01$ t-test are marked (See **Methods.** for details). **B.** More examples UMs and MMs matched to known motifs CTCF, SP1, PAX5 and TOPORS disrupted by somatic mutations show correlation with local methylation alteration.

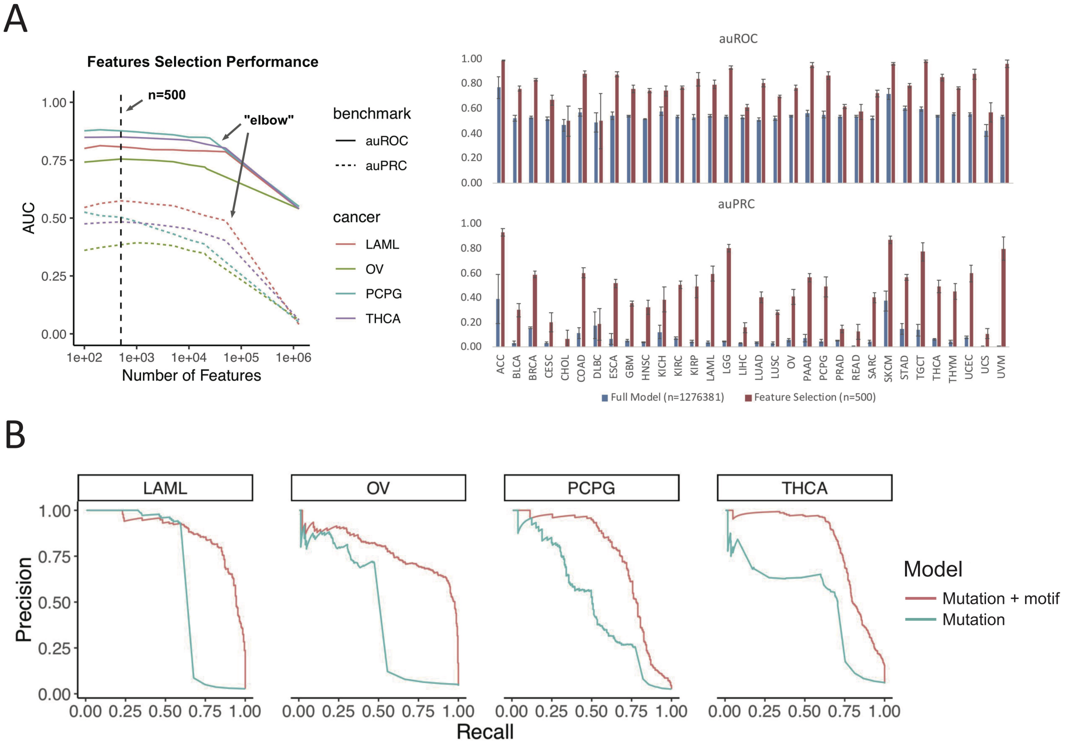


Figure 2.9: Motifs disrupted by mutations predict cancer subtype and survival. **A.** Feature selection reduces feature number while improving performance for the mutation-only diagnosis model. ”Elbow” indicates the number of features having positive feature importance scores. **B.** Performance evaluation of models predicting patient cancer subtype using auROC metric.

Chapter 3

Motto: Representing motifs in consensus sequences with minimum information loss

3.1 INTRODUCTION

Motif analysis is crucial for uncovering sequence patterns, such as protein-binding sites on nucleic acids, splicing sites, epigenetic modification markers and structural elements[168]. A motif is typically represented as a Position Weight Matrix (PWM), in which each entry shows the occurrence frequency of a certain type of nucleic acid at each position of the motif. PWMs are often visualized by sequence logo[168], which requires a graphical interface. Recently, several studies have shown the usefulness of representing motifs using kmers[169–172]; despite the power of this representation in machine learning models, it is cumbersome to have a set of kmers to characterize a single motif. In many scenarios, motifs can be sufficiently represented by regular expressions of the consensus sequences, such as [GC][AT]GATAAG[GAC] for the GATA2 motif. This representation is the most compact and intuitive way to delineate a motif. In the GATA2 motif example, the GATAAG consensus in the center is the most prominent pattern that would be read off the PWM or sequence logo. For this reason, consensus sequences are still widely used by the scientific community. Sequence pattern in the regular expression is used to highlight motif occurrence in popular genome browsers such as UCSC[173] and IGV[174]. Consensus sequences are assigned to *de novo* motifs and sequences for informative denotations[10, 11, 105, 175]. Wildcard-like sequence patterns are also supported in DNA oligo libraries synthesis by major vendors including Invitrogen, Sigma-Aldrich, and Thermo-Fisher.

However, current methods that convert PMWs to consensus sequence are often heuristic. One simple approach is taking the nucleotide with maximal frequency at each position to define the consensus sequence (eg. GGTCAAGGTCAC for ESRRB). Unsurprisingly, this could misrepresent positions with similar frequencies (eg. 0.26, 0.25, 0.25, 0.24, which should have been assigned as N). Alternatively, Douglas *et al.*[176] proposed in 1987 to follow a set of rules: use the single nucleotide with the highest frequency when it exceeds 0.50 and two times the second highest frequency; else, use the top two dinucleotides when their total frequencies exceed 0.75; else, use N. However, these rules are arbitrary, inflexible and lack a mathematical framework.

Here we present Motto, a sequence consensus representation of motifs based on information theory and ensures minimal information loss when converted from a PWM (**Figure 4.1**). We provide a standardized solution that determines the optimal regular expression of motif consensus sequence. We have also implemented an lightweight and easy-to-use Python package with versatile options for the biologists.

3.2 METHODS

Problem formulation

A positional weight matrix (PWM) maps $I \times P \rightarrow R$, where I is the set of indices of motif positions, and P is the frequency of the nucleotide in the motif. For a given position $i \in I$, let $C(i)$ denote the perceived frequencies for a combination of nucleotides, defined by equal frequencies shared among included nucleotides. For example, a $C(i)$ of [ACT] has the frequencies of [0.333, 0.333, 0, 0.333] for [A, C, G, T], respectively. Thus, we consider the optimal consensus sequence as a series of combination of nucleotides that has the most similarity between $C(i)$ and $P(i)$ for each position $i \in I$.

Minimal Jensen-Shannon Divergence (JSD) method

Here, we propose to use Jensen-Shannon Divergence (JSD) to measure the similarity between $C(i)$ and $P(i)$. JSD has been widely used in information theory to characterize the difference between distributions[177]. Using this metric, the combination of nucleotides with the least JSD from $C(i)$ to $P(i)$ will have the minimal "information loss", and is thus considered as the optimal consensus nucleotide.

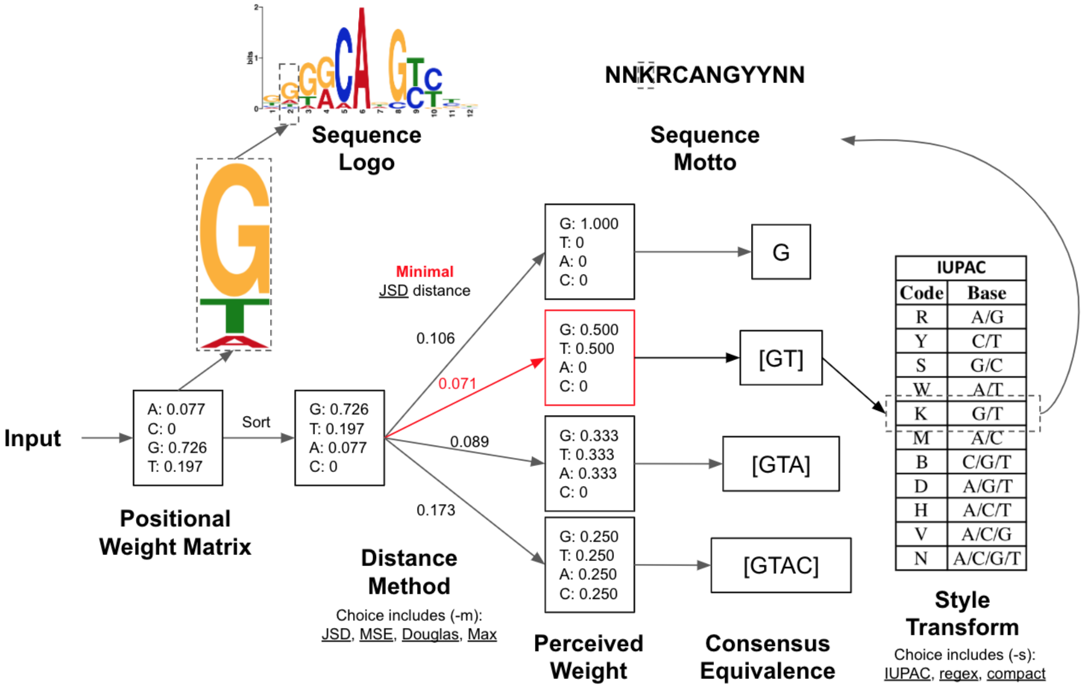
To efficiently compare JSD between all possible nucleotide combinations, we propose the following algorithm (**Figure 4.1**). Given a motif in its PWM form, having k positions, and n possible elements at each position, then the probability of jth element at ith position is given by $P(i, j)$, where $\sum_j P(i, j) = 1$, $i \in 1, 2, ..., k$, and $j \in 1, 2, ..., n$. First, we sort the elements of the PWM in descending order, so that:

$$P(i, j_1) \geq P(i, j_2) ... \geq P(i, j_n)$$

For example, at the 2nd ($i = 2$) position of the human transcription factor P73 (**Figure 4.1**), the nucleotides are sorted by occurrence frequencies so that:

$$P(2, "G") = 0.726 \geq P(2, "T") = 0.197 \geq P(2, "A") = 0.077 \geq P(2, "C") = 0$$

Next, we denote m as the number of different elements to be presented in the output consensus at the ith position, $m \in 1, 2, ..., n$. If a nucleotide is contained in a position in the consensus sequence, all the other nucleotides with higher frequencies at the ith position must



IUPAC	
Code	Base
R	A/G
Y	C/T
S	G/C
W	A/T
K	G/T
M	A/C
B	C/G/T
D	A/G/T
H	A/C/T
V	A/C/G
N	A/C/G/T

Figure 3.1: Overview of sequence Motto and comparison with sequence logo. Given a motif PWM as the input, Motto outputs a consensus that minimizes information loss. Here we show how the sequence Motto of the human transcription factor P73 is determined through the minimal JSD method.

also be included. Therefore, identification of the optimal consensus sequence is equivalent to identifying the optimal m.

When represented by the consensus sequence Motto, each nucleotide is considered as equally probable at a given position. For a position with $j_{i,1}, j_{i,2}, ..., j_{i,m}$ nucleotides, $C(i,1) = C(i,2) = ... = C(i,m) = 1/m$. The closer this distribution is to the original distribution of nucleotide frequencies, the better approximation of the consensus motif is to the original PWM. The optimal m (denoted as m^*) can be determined by minimizing the JSD between the two distributions:

$$m^* = argmin_m JSD(C(i,m), P(i) + q^2 \times m)$$

$$JSD(A,B) = \frac{1}{2}KLD(A,M) + \frac{1}{2}KLD(B,M)$$

$$M = \frac{1}{2}(A+B)$$

$$KLD(A,B) = \sum_{i=1}^{n} log(\frac{A(i)}{B(i)})$$

Here, $q \in [0,1]$ is the ambiguity penalty, a parameter input from the user to penalize large value of m in the output. When q=0, the optimal m^* marks the canonical minimal JSD. When q=1, $m^*$is guaranteed to be 1, hence the output consensus nucleotide is j_1(equivalent to using nucleotide with the maximal frequency). Thus, the optimal consensus nucleotide at the ith position is:

$$j_{i,1}, j_{i,2}, ..., j_{i,m^*}$$

Repeat this procedure for every position $i \in 1,2,...,k$, the final optimal consensus sequence is given by:

$$\{j_{1,1}, j_{1,2}, ..., j_{1,m^*}\}\{j_{2,1}, j_{2,2}, ..., j_{2,m^*}\} \ldots \{j_{k,1}, j_{k,2}, ..., j_{k,m^*}\}$$

Minimal Mean Squared Error (MSE) method

For comparison purposes, we have also implemented minimal mean squared error (MSE) method, which is another widely-used metric to measure distribution discrepancy[178]. The rest of the implementation is unchanged, except for the optimal m (m^*) is now determined by minimizing the MSE between the two distributions:

$$m^* = argmin_m MSE(C(i,m), P(i) + q^2 \times m)$$

$$MSE(A,B) = \frac{1}{n}\sum_{i=1}^{n}(A(i) - B(i))^2$$

Evaluating motif occurrence sites

We have collected 1156 common transcription factors from human and mouse from the databases of Transfac[179], Jaspar[180], Uniprobe[181], hPDI[182], and HOCOMOCO[104]. Each PWM is converted into consensus sequences, using default options of the four discussed methods: JSD (described above), MSE (described above), Douglas[176] and the naive approach of using the maximal frequency. Motif occurrence sites are determined in the human genome (hg19), matched by their regular expressions. The ground truth of the occurrence sites is determined by scanning the original PWMs with FIMO[107] using a 1e-5 *p*-value cutoff. The resulting *p*-values are converted into a significance score (-log(*p*-value)) and assigned to the matched motif occurrence sites from sequence Mottos. Thus, the area under the precision-recall curves[183](auPRC) is calculated by comparing the motif occurrence sites and their significance scores. Resulting auPRCs are averaged and a paired (by each motif) t-test is conducted to determine performance. Comparisons with significance (*p*-value ¡ 0.01) are shown (**Figure 4.3**).

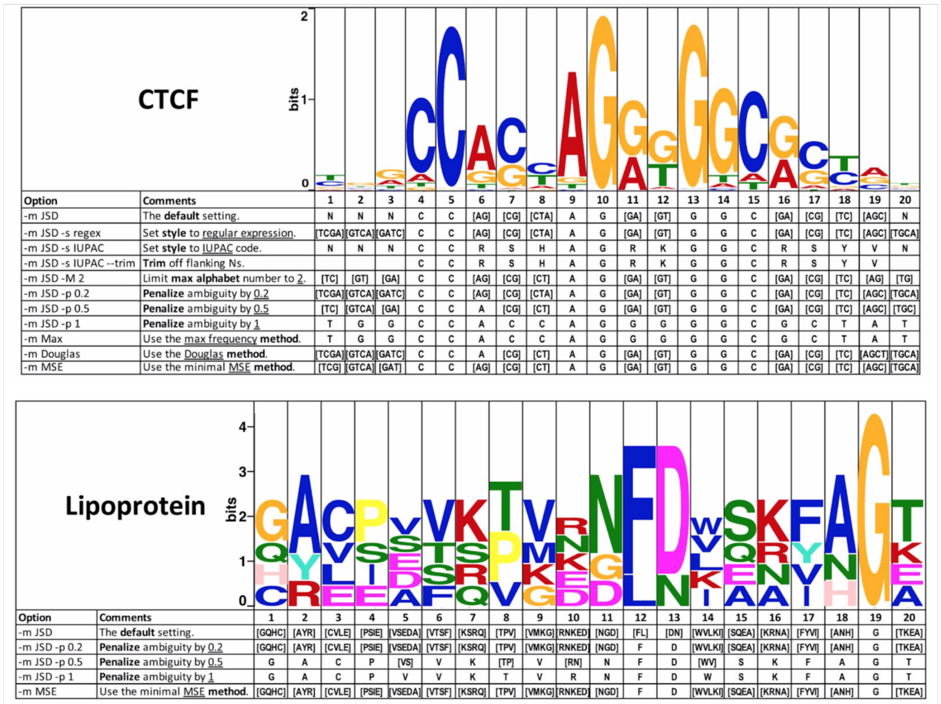
3.3 FEATURES AND EXAMPLES

Motto takes the MEME format of PWM as the input because of its popularity. The MEME format is supported by the majority of the motif databases[104], and the MEME suite provides packages for integrative analysis and conversion from other motifs formats[105]. The recently proposed kmer-based motif models also support conversion to MEME format[169, 170]. Our package is lightweight and open-source. The algorithm is efficiently implemented in python and the conversion for 1000 motif sequences typically takes less than two seconds. In addition, perhaps expectedly, downstream analysis like matching motif occurrences using the regular expression of sequence Motto is much faster (about 5 seconds for a common PWM on a chromosome, implemented in-house with python) than a conventional PWM scanning (about 1 minute, scanned with FIMO[107]. By default, the Motto package takes a motif in the MEME format, parses the header to get the nucleotide, computes the optimal consensus sequences based on the minimal JSD method, and then outputs the sequence in a compact format (**Figure 4.1**). Motto provides flexibility at each step along this process. Input can be from a file, or from standard input, and Motto can consider nucleotides, amino acids and customized alphabets such as CpG and non-CpG methylation[69]. Four methods are made possible for comparison: maximal probability (Max), heuristic Douglas method (Douglas), minimal mean squared root (MSE) method, and our proposed minimal Jensen-Shannon Divergence (JSD) (See **Methods**). Three output styles are provided: (1) IUPAC uses a single alphabet to represent the combination of nucleotides (eg, S for [CG]) and is the most compact form, but require reference to the nomenclature[184]; (2) regular expression ("regex") enumerate all output consensus nucleotide ranked by occurrences and is recommended for downstream analysis, such as motif occurrence and oligo designs; (3) "compact" (the default) is the same as "regex", except that it replaces [ACGT] with N. To trim off Ns ([ACGT]s) at both ends of the output sequences, a optional flag "–trim" is provided. If the users prefer consensus with more certainty (eg, prefer [AC] to [ACG]), they can use either

”–maxAlphabet” as a hard limit to the number of alphabets allowed, or use ”–penalty” to penalize ambiguity (see **Methods**).

Effects of these options are shown using an example of human transcription factor CTCF (**Figure 4.2, upper panel**). Unsurprisingly, MSE, JSD, and Douglas are more representative than the naive maximal probability methods. For example, position 1,2,3 and 20 with low information content (¡0.2) in CTCF, are justifiably called as *N* by JSD and Douglas, which is an improvement over strictly calling the top nucleotide. MSE considers [TCG] and [GAT] more representative at the 1st and the 3rd position but agrees with JSD and Douglas at the 2nd and 20th. Similarly, JSD, MSE, and Douglas successfully capture strong double-consensus patterns at indices 7, 11, 12 and 16, which maximal probability fails to capture. The advantage of JSD over Douglas is noticeable at index 6, where the logo of CTCF shows a dominating AG consensus. While JSD finds this co-consensus, Douglas disregards G that barely misses the cutoff. In addition, at index 19, the logo of CTCF shows a strong three-way split among A, G and C, but Douglas, by its rules (as described previously), ignores all such triple patterns. In addition, among the four methods, only the JSD and MSE are capable of generating consensus sequences for amino acid motifs[185] (**Figure 4.2, lower panel**). Due to its arbitrary nature, heuristic methods like Douglas have difficulties defining decision boundary for motifs more than 4 nucleotides. In such cases, JSD and MSE provide more mathematically rigorous information than Douglas and oversimplified maximal consensus methods. With increased penalty level at 0, 0.2, 0.5 and 1 respectively, the consensus sequence smoothly progresses towards single nucleotide consensus (**Figure 4.2**). Such flexibility gives an advantage to users that are biased towards more defined consensus results.

To quantify how well these four methods summarize the information in the original PWMs, we converted 1156 common human and mouse transcription factors to consensus sequences and compared their matched occurrences (by regular expression) in the human genome (hg19) with conventional motif sites scanned by FIMO[184] with PWMs, which is how conventionally motif sites are determined (see **Methods**). We observe that using the JSD method has resulted in



Option	Comments	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
-m JSD	The **default** setting.	N	N	N	C	C	[AG]	[CG]	[CTA]	A	G	[GA]	[GT]	G	G	C	[GA]	[CG]	[TC]	[AGC]	N
-m JSD -s regex	Set **style** to regular expression.	[TCGA]	[GTCA]	[GATC]	C	C	[AG]	[CG]	[CTA]	A	G	[GA]	[GT]	G	G	C	[GA]	[CG]	[TC]	[AGC]	[TGCA]
-m JSD -s IUPAC	Set **style** to IUPAC code.	N	N	N	C	C	R	S	H	A	G	R	K	G	G	C	R	S	Y	V	N
-m JSD -s IUPAC --trim	**Trim** off flanking Ns.				C	C	R	S	H	A	G	R	K	G	G	C	R	S	Y	V	
-m JSD -M 2	Limit **max alphabet** number to 2.	[TC]	[GT]	[GA]	C	C	[AG]	[CG]	[CT]	A	G	[GA]	[GT]	G	G	C	[GA]	[CG]	[TC]	[AG]	[TG]
-m JSD -p 0.2	**Penalize** ambiguity by 0.2	[TCGA]	[GTCA]	[GATC]	C	C	[AG]	[CG]	[CTA]	A	G	[GA]	[GT]	G	G	C	[GA]	[CG]	[TC]	[AGC]	[TGCA]
-m JSD -p 0.5	**Penalize** ambiguity by 0.5	[TC]	[GTCA]	[GA]	C	C	A	[CG]	[CT]	A	G	[GA]	[GT]	G	G	C	[GA]	[CG]	[TC]	[AGC]	[TGC]
-m JSD -p 1	**Penalize** ambiguity by 1	T	G	G	C	C	A	C	C	A	G	G	G	G	G	C	G	C	T	A	T
-m Max	Use the max frequency **method**.	T	G	G	C	C	A	C	C	A	G	G	G	G	G	C	G	C	T	A	T
-m Douglas	Use the Douglas **method**.	[TCGA]	[GTCA]	[GATC]	C	C	A	[CG]	[CT]	A	G	[GA]	[GT]	G	G	C	[GA]	[CG]	[TC]	[AGCT]	[TGCA]
-m MSE	Use the minimal MSE **method**.	[TCG]	[GTCA]	[GAT]	C	C	[AG]	[CG]	[CT]	A	G	[GA]	[GT]	G	G	C	[GA]	[CG]	[TC]	[AGC]	[TGCA]

Option	Comments	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
-m JSD	The **default** setting.	[GQHC]	[AYR]	[CVLE]	[PSIE]	[VSEDA]	[VTSF]	[KSRQ]	[TPV]	[VMKG]	[RNKED]	[NGD]	[FL]	[DN]	[WVLKI]	[SQEA]	[KRNA]	[FYVI]	[ANH]	G	[TKEA]
-m JSD -p 0.2	**Penalize** ambiguity by 0.2	[GQHC]	[AYR]	[CVLE]	[PSIE]	[VSEDA]	[VTSF]	[KSRQ]	[TPV]	[VMKG]	[RNKED]	[NGD]	F	D	[WVLKI]	[SQEA]	[KRNA]	[FYVI]	[ANH]	G	[TKEA]
-m JSD -p 0.5	**Penalize** ambiguity by 0.5	G	A	C	P	[VS]	V	K	[TP]	V	[RN]	N	F	D	[WV]	S	K	F	A	G	T
-m JSD -p 1	**Penalize** ambiguity by 1	G	A	C	P	V	V	K	T	V	R	N	F	D	W	S	K	F	A	G	T
-m MSE	Use the minimal MSE **method**.	[GQHC]	[AYR]	[CVLE]	[PSIE]	[VSEDA]	[VTSF]	[KSRQ]	[TPV]	[VMKG]	[RNKED]	[NGD]	F	D	[WVLKI]	[SQEA]	[KRNA]	[FYVI]	[ANH]	G	[TKEA]

Figure 3.2: Example usage using human CTCF (**upper panel**) and lipoprotein binding sites from Bailey 1994[185] (**lower panel**). The original PWM is shown in sequence logo. Different Motto options resulted in various consensus sequence output at each position. In particular, ”-m/–method” specifies the method: JSD (default), MSE (minimal mean square error), Douglas[176], or Max (using maximal frequency at each position); ”-s/–style” specifies the output style: IUPAC[184] (single alphabet for nucleotide combinations), regex (regular expression), or compact (convert [ACGT] to N in regex); ”-t/–trim” is an option for trimming off the flanking Ns; ”-p/–penalty” specifies a weight between 0 to 1 that penalizes ambiguity at each position. (For details see **Methods**)

the best (0.81± 0.01) area under the Precision-Recall curve (auPRC), significantly (*p*-value ¡ 0.01) when compared with existing alternative methods, including MSE (0.76 ± 0.01), Douglas (0.76 ± 0.01), and maximal frequency (0.53 ± 0.04) (**Figure 4.3**).

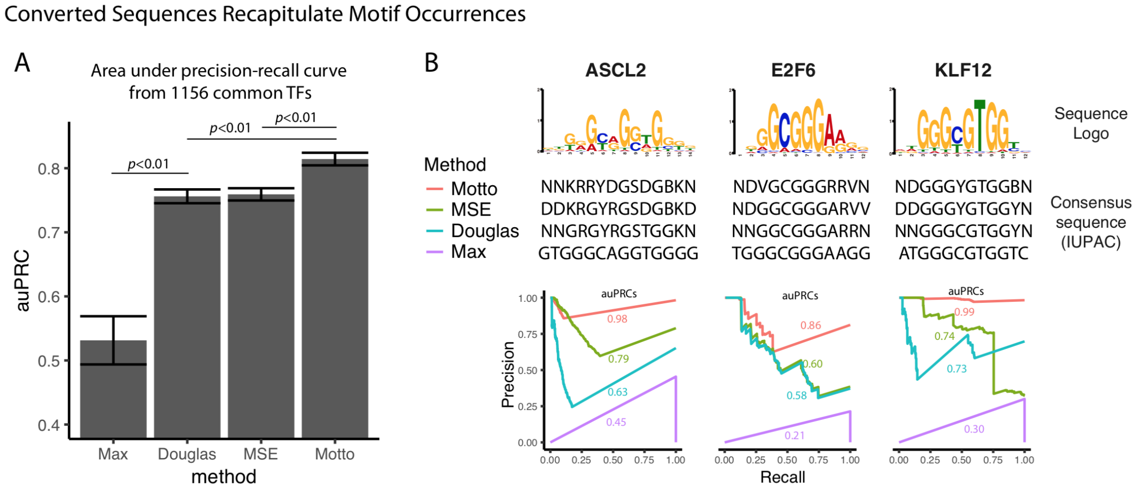


Figure 3.3: Converted sequence Mottos recapitulate motif occurrence sites of 1156 common human and mouse transcription factors (TFs) in the human genome (hg19). **A**. Averaged area under the precision-recall curve (auPRC) using Motto (default method with minimal JSD) compared with existing alternative methods. *P*-value determined by paired t-test. **B**. Comparison in three example TFs showing the differences of consensus sequences (shown in IUPAC[184] coding for better alignment) and performances.

In summary, Motto provides a mathematical framework and a set of convenient features to epitomize PWMs in a compact, intuitive and informative manner.

Chapter 3, in full, is a reprint of the material as it appears in ”Motto: Representing motifs in consensus sequences with minimum information loss. Mengchi Wang, David Wang, Kai Zhang, Vu Ngo, Shicai Fan, Wei Wang.” in Biorxiv, 2019. The dissertation author was a primary investigator and author of this paper.

Chapter 4

CRISPY: a pipeline that unifies various CRISPR/Cas9 functional screening protocols and leverages existing genetic and epigenetic knowledge

4.1 INTRODUCTION

The function of genes is determined by an intricate network that involves both local genetic elements and cell-specific environment. This includes both cis-acting genetic elements, such as promoter and enhancers, as well as trans-acting elements such as transcription factors. Recent advancement by large-scale studies, such as ENCODE[65], ROADMAP[66], and TCGA[67], has provided a comprehensive atlas for genetically function elements, particularly in the non-coding regions. For example, the ENCODE consortium has provided a web service named SCREEN (Search Candidate cis-Regulatory Elements by ENCODE(http://screen.encodeproject.org), which output tissue-specific cis-regulatory elements candidates based on existing cellular signatures from H3K27ac, H3Kme1, H3Kme3, DHS, and ATAC-seq data. However, the functionality of the putative candidates needs to be further determined experimentally.

Currently, there are multiple assays to characterize enhancer-promoter activity. For example, MPRA (Massive Parallel Reporter Assays)[186] and STARR-seq (Self-Transcribing Active Regulatory Region Sequencing)[187] combine traditional reporter assays with RNA-seq technology to profile enhancer activity. However, there are several limitations. For example, MPRA is limited by the size ($\sim 200bp$) and number (currently $< 100,000$) of oligonucleotides. On the other hand, in the STARR-seq approach, the DNA fragments are cloned and taken out of the native location, potentially causing artifacts[188]

In comparison, the recently popularized CRISPR/Cas9 mediated genetic screening has been increasingly recognized as the ”gold standard” for screening functional elements of target genes[189]. Indeed, CRISPR/Cas9 screening has addressed these aforementioned limitations by providing the extended capability of oligonucleotide design. More importantly, it enables the interrogating genome function for both transcribed genes and non-coding regions in their normal genetic and epigenetic context[189]. Currently, there are two distinct yet complementary protocols for functional element screening. The first strategy is using targeted single-sgRNA CRISPR/Cas9 (”single-sgRNA” for short) mediated genetic screening. This protocol targets pre-defined putative cis-regulatory candidates, typically derived from the chromatin and histone modification, providing an unparalleled resolution into the targeted regions[190]. For example, Diao et al in 2016 interrogated 174 candidate regulatory sequences within the 1-Mbp POU5F1 locus in human embryonic stem cells (hESCs) and discovered enhancers and new mechanisms at the non-coding regions[190]. On the other hand, Diao et al.[191] in 2017 proposed another method termed ”CREST-seq” (cis-regulatory element scan by tiling-deletion and sequencing). This method is unbiased and systematically delete cis-regulatory elements in a tiling fashion, providing insights into previously under-reported elements. As a result, 45 cis-regulatory elements have been identified in a 2-Mb POU5F1 locus in human embryonic stem cells, with some located at regions that would otherwise be excluded from pre-defined putative candidates[191].

Despite the fast advancement in CRISPR/Cas9 mediated screening, a unified, easy-to-use pipeline that accommodates both single-guide CRISPR and CREST-seq is lacking. Another challenge of current CRISPR screening pipeline is the lack of reference for predicted functional elements for downstream experimental validation. Here we present CRISPY: a lightweight, versatile, customizable pipeline that streamlined CRISPR screening from both single-guide CRISPR and CREST-seq experiments. CRISPY provides several key advantages compared to current pipelines. First, it takes advantage of robust algorithms like alpha-RRA[192] and flexible user interface to allow confident peak calling in either the targeted regions or unbiased genomic segments. Further, CRISPY makes a comparison between multiple experiments accessible and presents key visualization for quality control purposes. Finally, CRISPY provide corroborating evidence of the peak candidates from screening results by integrating prior knowledge from data such as histone modifications, DHS, and ATAC-seq in existing human and mouse tissues and cell lines. By combining the prior knowledge with a random forest model, we show that CRISPY successfully identified previously unreported functional elements for target genes including Sox2, Dppa2, and FMR1. In order to test the function of the called peaks, elements in Dppa2 are further validated experimentally through CRISPR-knockdown and CRISPR-i. We show that this novel strategy which combines unified CRISPR/Cas9 protocols and existing genetic and epigenetic knowledge from multiple samples can provide improved performance in the discovery of functional elements. The pipeline

4.2 METHODS

4.2.1 Preprocessing inputs

To unify the input from various methods, we require the input to provide three key information invariably presented in all CRISPR/Cas9 protocols. The region file contains the targeted genomic regions in the ”∗.bed” format. This file can be customized and will be the basic unit where peak calling takes place. For example, in a single-sgRNA protocol, the target region file can specify the putative active candidates, usually determined by ATAC-seq, H3K27ac, and CTCF binding sites[190]. The read file requires is the mapped read counts of each sgRNA across multiple experiments. We require the user to specify the experiments that should be included in either foreground or background. On the other hand, reads are usually processed by aligners such as BWA[193] or BOWTIE[194, 195] and have been separated from the CRISPY pipeline to maintain its lightweight and modularity. Finally, we require the information of the sgRNAs, including their genomic location, and their label in the group (for example, test, positive control, or negative control). To unify the input files from various platforms(Windows/macOS/Linux), CRISPY remove the special white spaces from the input. Finally, CRISPY automatically checks and prompts installation (if necessary) all its dependencies, which are the bedtools[196](for basic ”.bed” file manipulation), edgeR[197](for the negative binomial test), and ggplot2(for visualization).

4.2.2 Quality control

Quality control steps are organized in two phases: before the determination of the significance of sgRNA enrichment (ie., peak calling in the next section), and after. In the first phase, to evaluate the raw mapping coverage, we plot the sgRNA read distribution across provided samples in a violin-boxplot (**Figure 4.5A**). Principle component analyses are conducted to show the similarity among experiments, with labels specified in the input files described earlier (**Figure 4.5B**). To ensure the quality of the reads, we keep sgRNAs that have ≥ 5 reads in at least 50% of the foreground (or background) experiments. For example, if there are 5 foreground samples and 2 background controls, a sgRNA needs to have ≥ 5 reads in at least 3 (50% of 5 rounded up) or 1 (50% of 2) background samples. These parameters can be adjusted in the pipeline. As a result, the remaining sgRNA with sufficient coverage have their reads plotted in a cumulative frequency plot (**Figure 4.5C**). To compare the sgRNAs in the foreground and background, we plot the averaged read counts of each sgRNAs on a scatter plot (**Figure 4.5D**).

To ensure a robust result and remove the batch effect, we design an optional quantile normalization with customization, which allows users to specify how to normalize read counts of the sgRNA among samples. For example, -q ”FG1, FG2; BG1, BG2, BG3” will equalize the quantile distribution among FG1 and FG2 samples, then among BG1-3 samples. Following previous studies[198], read counts of the sgRNAs are further normalized by the TMM[198] (Trimmed Mean of M-values) methods from the edgeR package[197]. Briefly, TMM equalizes the mean of M-values between experiments, which equates the overall expression levels of the genes between samples, under the assumption that the majority of them are not differentially expressed. Both steps remove the artifacts from systematic bias and improve the robustness of the results.

After peak-calling, the quality of the sgRNAs is profiled in an MA plot (total read count by fold change) and a significance scatter plot (fold change by p-value) (**Figure 4.5**). To guide the users to choose the appropriate parameters for peak calling, we provide a guide table that calculates an ”external FDR” defined by the of negative sgRNAs (specified in the sgRNA file described earlier) that would be included, if a p-value cutoff (by the -n option) is given in the peak-calling step (**Table S1**).

Peak-calling

Peak-calling is performed broadly following previous studies[190, 191], using edgeR[198] and RRA[199](robust rank aggregation). Modifications are provided to accommodate various experimental protocols. Briefly, we estimate the statistical significance (using a negative binomial test) of each sgRNA in foreground samples in contrast to background samples using the edgeR package. Next, we converted the resulting p-values to -log(P) values of the sgRNAs and show in the ”$*$ .sgRNA.bedgraph” files for visualization. Next, we used the target genomic region file (described in the previous section) provided by the users to identify the partition of the genome. For example, for ”single-sgRNA” experiment, the targeting regions would be predefined genomic segments containing ATAC-seq, H3K27ac, and CTCF binding sites[190]. In comparison, previous studies have used non-overlapping bins of arbitrary length (eg. [191] has used 50bp), we provide an option to partition the genomic regions by sliding from one sgRNA to the next. This strategy ensures the highest resolution possible given the design of the oligo library and equates using 1bp sliding windows with much faster processing speed(data not shown). To identify the positive ”bin” of the target regions, we perform the RRA step described previously[192]. Briefly, the sgRNAs passing a user input significance level have the logP values converted to rankings, and a bin with more high ranking sgRNAs will be identified as significant. Finally, the p-values of each bin are adjusted to FDR by the Benjamin-Hochberg procedure. Results are visualized in ”$*$ .region.bedgraph”. We note that sometimes when a region has lower coverage, the RRA test would underestimate the significance of a bin with a few but highly enriched sgRNAs. To address this, we have also provided a high-resolution signal track based on average fold change. In summary, we follow the same steps until RRA and average the fold change of the sgRNAs in the same bin in ”$*$ .fc.bedgraph” for further visualization.

4.2.3 Random forest model

To boost the confidence of the identified positive peaks, we build a random-forest model using existing data. First, we download the H3K27ac from the ENCODE consortium in mm10, where MACS2[200] has been used to call peaks by the ENCODE DCC[65]. We also download the RNA-seq from the ENCODE consortium. In total, 30 pairs of H3K27ac and RNA-seq from various tissues and cells have been collected. Next, an ”activity score” is given to each positive peak identified from the peak-calling step, by using the averaged H3K27ac (by default, can be expanded to other datasets, such as HiC or ATAC-seq) peaks MACS2 score in the 1000bp (by default, can be changed). Next, we use this activity score as the feature to predict the expression level (TPM) of the target gene (In this case, Sox2, Dppa2, or FMR1) from the 30 cell lines and tissues. Next, to build a machine learning model, we ran the random forest using feature trees $n = 500$, with default parameters from the R package ”randomForest”[201]. The ”importance score” of each peak is determined by the out-of-bag Gini-index from random-forest and visualized from the ”$*$ .imp.bedgraph”.

4.3 RESULTS

4.3.1 CRISPY identified cis-regulatory elements of Dppa2

To evaluate the cis-regulatory elements of Dppa2, we performed both single-sgRNA and CREST-seq CRISPR-cas9 screening following previous protocols[190, 191](**Figure 4.1**).Briefly, we introduced a large number of overlapping genomic deletions that are introduced to the 3Mb region ($chr16 : 46000000 - 49000000$) of Dppa2 in the mouse with paired sgRNAs. We collected the resulting cells with reduced expression of Dppa2, and establish them as the foreground while using the unsorted cells as the background. Next, we apply CRISPY to the sgRNA mapped from these cells using BOWTIE[194] as described (**Figure 4.1**, see **Methods.**).

To standardize the input from various CRISPY protocols, CRISPY requires three basic files, the input read counts of each sgRNA, the genomic position of the sgRNAs, and the target region of interest. For example, in single-guide experiments, users would input targeted open chromatin regions. In CREST-seq (or the default setting), CRISPY would automatically generate a single base-pair resolution based on the coverage of the input sgRNA oligo design (see **Methods**).

Out of the sgRNA pairs we have designed, we identified 43 CRISPY peaks from CREST-seq experiments, and 34 single-sgRNA CRISPY peaks, using the FDR cutoff of 0.05 (**Figure 4.2**). The average size of the peaks is $\sim 2kb$ for the CREST-seq and $\sim 1kb$ for single-sgRNA screening. Out of called CRISPY peaks, we observed strong enrichment ($> 2fold$) for H3K27ac, H3K4me1, CTCF, and ATAC-seq downloaded from the ENCODE in mESC[65].

Interestingly, we noticed that the biological data output from CRISPR experiments can

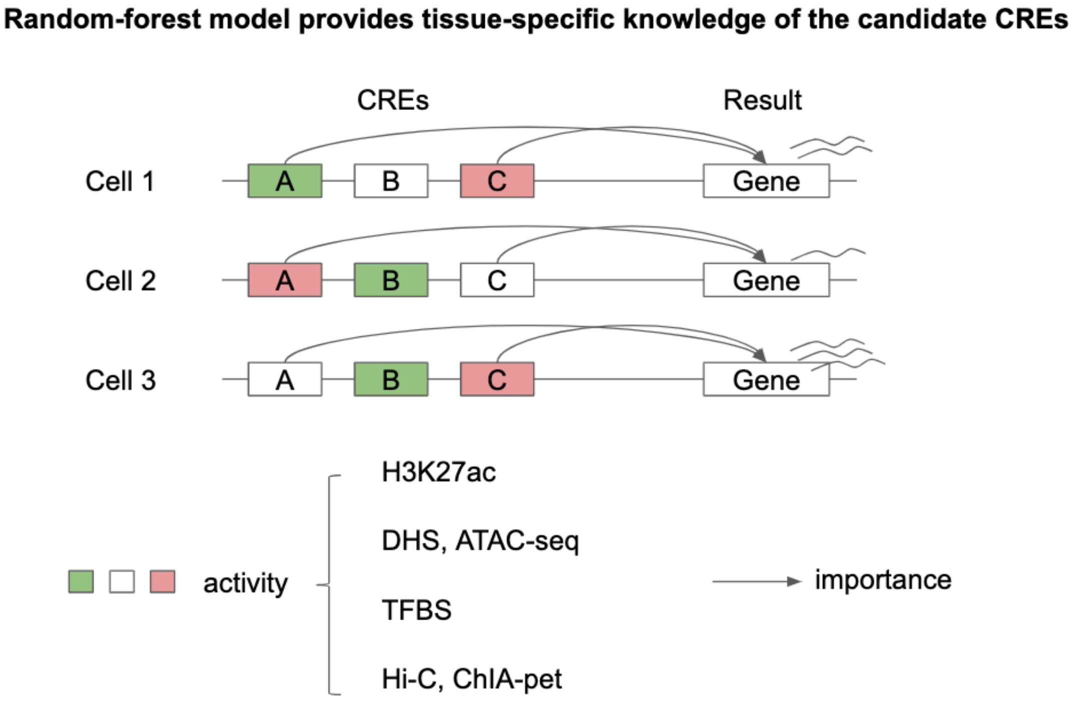


Figure 4.1: CRISPY pipeline Overview

be complicated. The data from a reliable and consistent sample is crucial to the downstream computational analysis. This reality necessitates stringent yet intuitive quality control. CRISPY provides visualization of the distribution of the reads across samples, and PCA to detect the heterogeneity among experiments. In this real biological example (**Figure 4.5**), we show that treatment sample #1 (T1) has a bimodal distribution when compared to T2 and T3, while the PCA plot shows T1 is far from T2 and T3. These suggest T1 is different from T2 and T3 in the total number of reads, and on the distribution of the sgRNA reads. Similarly, control sample #1 (C1) is marked by higher reads from C2, C3, and C4, which clustered together in the PCA plots, suggesting C1 is inconsistent with the rest of the control samples. Therefore, it is highly recommended to remove T1 and C1 from foreground and background, respectively.

Tracks show the significance for each output files, in the order of sgRNA logP value, RRA FDR value, and called CRISPY peaks. Validation is carried out by dual CRISPR knock out. Annotation files have been downloaded from the ENCODE data portal with relevant peaks from

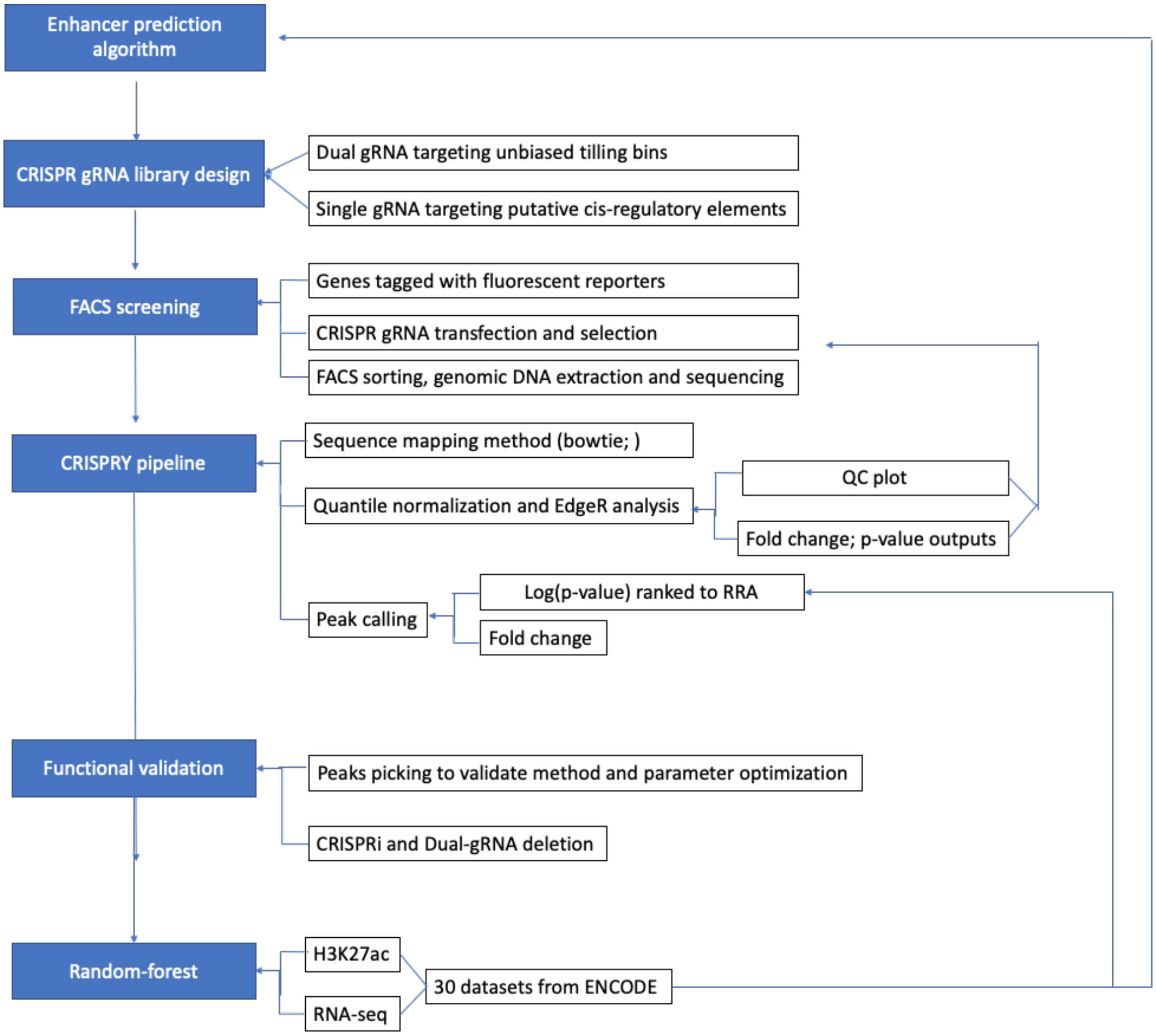

Enhancer prediction algorithm
CRISPR gRNA library design
Dual gRNA targeting unbiased tilling bins
Single gRNA targeting putative cis-regulatory elements
FACS screening
Genes tagged with fluorescent reporters
CRISPR gRNA transfection and selection
FACS sorting, genomic DNA extraction and sequencing
CRISPRY pipeline
Sequence mapping method (bowtie;)
Quantile normalization and EdgeR analysis
QC plot
Fold change; p-value outputs
Peak calling
Log(p-value) ranked to RRA
Fold change
Functional validation
Peaks picking to validate method and parameter optimization
CRISPRi and Dual-gRNA deletion
Random-forest
H3K27ac
RNA-seq
30 datasets from ENCODE

Figure 4.2: CRISPY signal for the 3Mb region around Dppa2.

mESC, including ChromHMM annotation, H3K27ac, H3K4me1, H3K4me3, CTCF, RAD21, ATAC-seq, and CCRE determined by the ENCODE SCREEN[65]

4.3.2 Random-forest model using existing knowledge from various cell lines.

One challenge we face in CRISPR screening is verification of the called candidate peaks. Besides experimental methods, many computational methods been proposed to identify the regulatory elements (such as enhancers) that regulate the expression of target genes[202]. For example, Yun et al[203] have proposed to use a tensor-based algorithm that decomposite the epigenetics information to predict enhancer-promoter interactions in mESC. Other models leverage the correlation between enhancer and promoters, relying on regression or machine learning to interrogate the relationship between enhancer and promoters. Despite the variety of the models and features used to predict enhancer-promoter interactions, one major challenge for the computational models is the problem of overtraining. This is usually caused by the small sample size with which the initial association between the promoters and enhancers are established. Combined with a large number of candidates elements that exceeds sample size, this will derive confident predictions and leads to false discovery.

To address these limitations, we leverage the prior knowledge of epigenome and focus on using H3K27ac to predict the element for the target gene. The advantage is that with focused elements pre-filtered by CRISPR, we can have better resolution using a robust regression model. Therefore, our prediction will be unparalleled accuracy based on prior knowledge, as well as the cell-type-specific experimental validation from the CRISPR results. The tradeoff is that we will only have predictions for a handful of CRE elements for the target gene (ie., $\sim$ 20 target candidate for Dppa2 in this study)

In our random forest model, we associate two values for the CRE-gene interaction (**Figure 4.3**).(1) An inherent ”importance” score that characterizes the function of the CRE towards the target gene. This is the target value to learn from the random forest model. For example, an enhancer will have higher importance for its target gene, than an unrelated nearby region. (2) An ”activity” score used to assign the cell-type-specific weight on each CRE. For example, we use the average H3K27ac score nearby 1000 bp of each CRE to characterize how much ”potential” of the importance of the CRE is realized. In other words, this differentiates an active enhancer from an inactive enhancer. As a result, we have assigned importance scores to 15 CRE candidates identified from the CREST-seq experiment, with OOB Gini importance *score* $>$ 0.015 (**Table 4.1**).

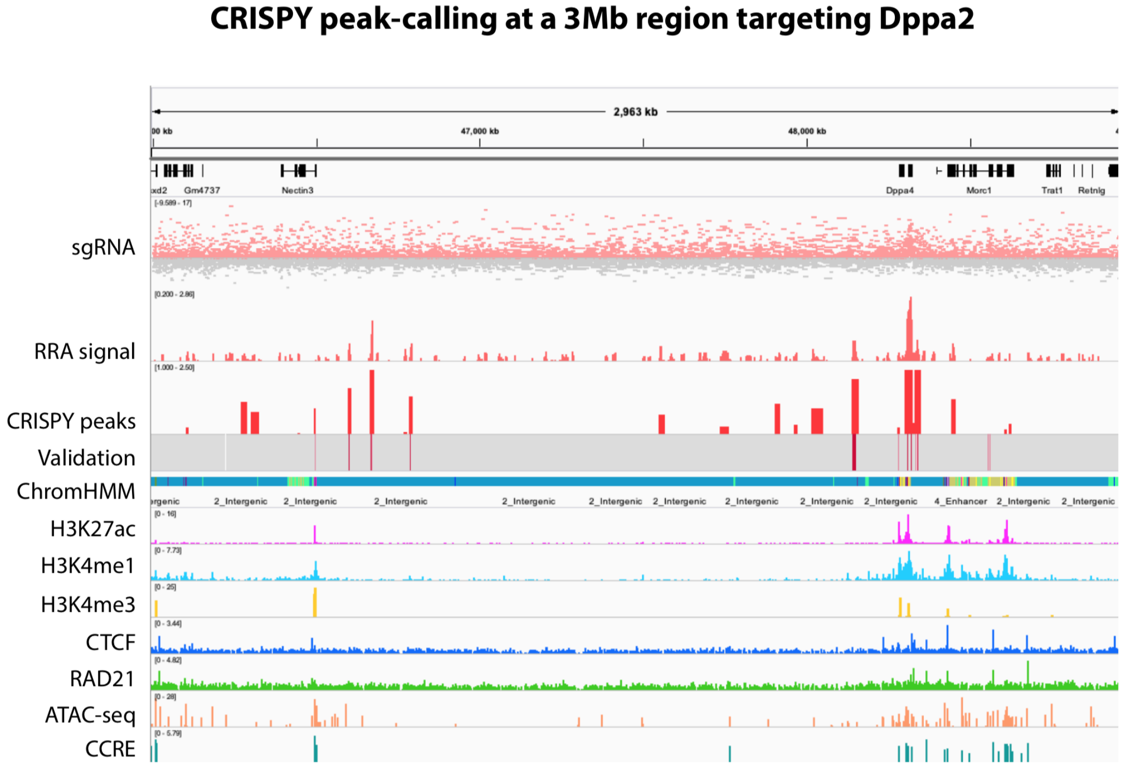

CRISPY peak-calling at a 3Mb region targeting Dppa2

Figure 4.3: random-forest model provides tissue-specific knowledge of the candidate CRES by integrating existing genetic and epigenetic data.

4.3.3 Experimental Validation

To validate the candidate CREs combinatorially selected from CRISPY peaks and the random-forest model, we conducted CRISPR-Cas9 dual knockout experiment (**Figure 4.2,4.4**). As a result, we observe 8 out of the 11 selected CRE have shown significant ($p < 0.05$) reduced expression of Dppa2.

Chapter 4, in full, is currently being prepared for submission for publication of the material as it would appear as ”CRISPY: a versatile pipeline for CRISPR functional screening. Mengchi Wang, Xiaoyu Yang, Guoqiang Li, Xingjie Ren, Yin Shen, Bing Ren, Wei Wang” The dissertation author was a primary investigator and author of this paper.

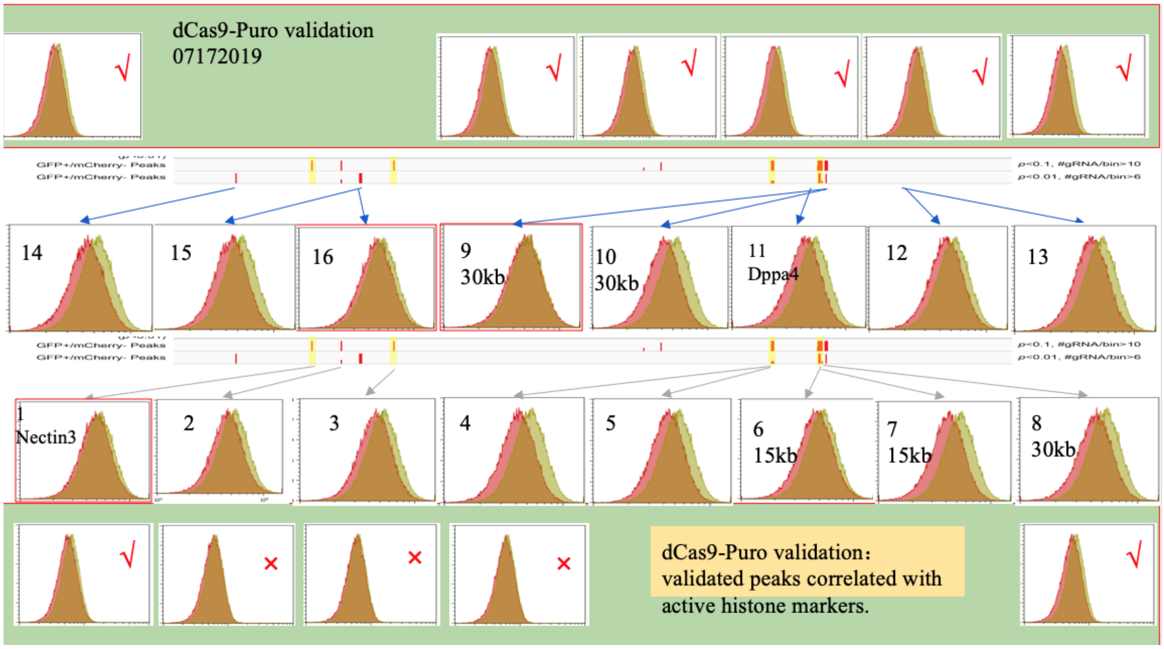
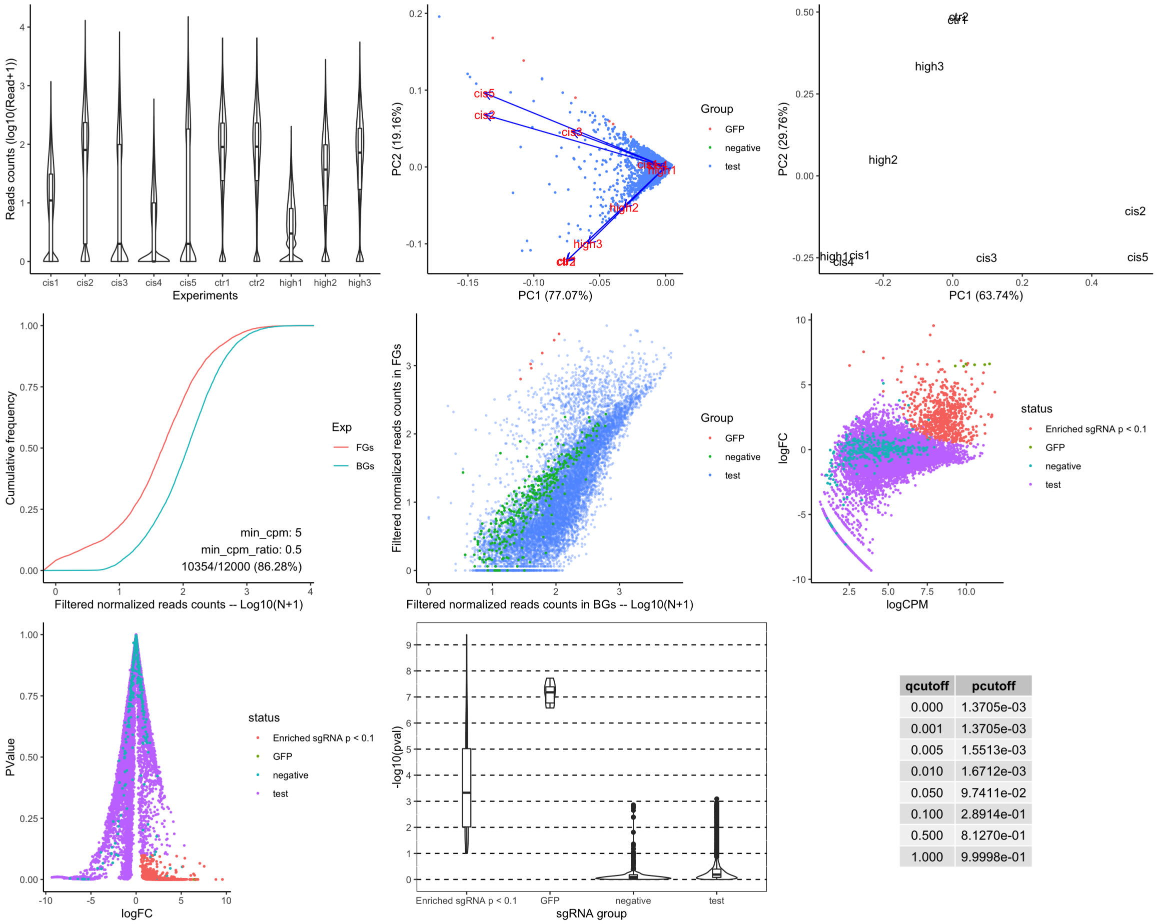


Figure 4.4: Experimental validation of CRISPR peaks by dual CRISPR knockout.

4.3.4 AVAILABILITY

CRISPY is available at https://github.com/MichaelMW/crispy

4.4 SUPPLEMENTARY



qcutoff	pcutoff
0.000	1.3705e-03
0.001	1.3705e-03
0.005	1.5513e-03
0.010	1.6712e-03
0.050	9.7411e-02
0.100	2.8914e-01
0.500	8.1270e-01
1.000	9.9998e-01

Figure 4.5: Quality control of sgRNA. **A**. distribution of the total read counts of each sgRNAs by experiments. **B**.PCA of experiments using sgRNA reads as features. **C.** The cumulative reads of sgRNAs from the foreground (FG) and background(BG). **D.** The scatterplot for the normalized read counts of the sgRNAs in foreground and background.

Table 4.1:

CRE location	OOB importance
*chr*3_34650423_34650923	0.0344
*chr*3_35224583_35225083	0.0265
*chr*3_35754047_35754547	0.0261
*chr*3_34664633_34665133	0.0241
*chr*3_34648963_34649463	0.0211
*chr*3_34649781_34650281	0.0206
*chr*3_34019705_34020205	0.0186
*chr*3_36552063_36552563	0.0180
*chr*3_34650797_34651297	0.0170
*chr*3_35405270_35405770	0.0168
*chr*3_34653503_34654003	0.0168
*chr*3_34648663_34649163	0.0162
*chr*3_35324061_35324561	0.0159
*chr*3_36066007_36066507	0.0159

Chapter 5

Concluding remarks

Current research for liquid biopsy is benefiting from two contributing factors: the increasing sequencing power that follows Moore's law[204], and the booming patient studies linking their molecular profile to early cancer and treatment responses[80, 205, 206]. Therefore, while currently available assays have low-dimension features (i.e., 10-100 strong biomarkers) due to limited variant data[207], we predict future studies to have high-dimension predictors. As we can see from the evolution of prediction models, we also predict the adoption of deep learning models that thrives on the plethora of datasets. Both trends require a deeper understanding of the interplay between existing features (i.e., DNA methylation and DNA variant) to unlock new predictive features.

To distinguish between cancer and normal cfDNA, the major challenge is the limited number of recurring biomarkers, and how to detect them frugally given the finite detection limit. Recent advances in technology have made possible the simultaneous detection of both DNA variants and methylation variants on cfDNA[95]. As a result, we would argue that the most cost-effective strategy is to adopt the prior knowledge of how DNA and methylation interact with each other to enable higher confidence and sensitivity. In addition, we have recently discovered DNA motifs that regulate histone modifications[11, 71] and showed that the altered DNA motif leads to abolished histone modification, which is also associative in cancer[208]. These *cis-acting* motifs can be leveraged to reveal information on the state of histones, which is not readily available in cfDNA[75]. Further, DNA patterns are also important in establishing local secondary structures, which has been reported as an epigenetic determinant of cancer genome[209]. Clark et al. have reported a sequence pattern in the secondary structures as a hotspot for DNA methylation in human breast cancer patients[40].

Taken together, we believe the ever-growing research revealing genetic-epigenetic interplay has opened doors to previously underrepresented strategies in biomarker selection, and points to new perspectives in characterizing DNA variants in combination with epigenetic signatures.

Bibliography

1. Razin, A. & Cedar, H. DNA methylation and gene expression. en. *Microbiol. Rev.* **55,** 451–458 (Sept. 1991).

2. 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., Gnirke, A. & Meissner, A. Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500,** 477–481. ISSN: 1476-4687 (2013).

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

4. Rose, N. R. & Klose, R. J. Understanding the relationship between DNA methylation and histone lysine methylation. en. *Biochim. Biophys. Acta* **1839,** 1362–1372 (Dec. 2014).

5. Rasmussen, K. D. & Helin, K. Role of TET enzymes in DNA methylation, development, and cancer. en. *Genes Dev.* **30,** 733–750 (Apr. 2016).

6. Blattler, A. & Farnham, P. J. Cross-talk between site-specific transcription factors and DNA methylation states. en. *J. Biol. Chem.* **288,** 34287–34294 (Nov. 2013).

7. Ravichandran, M., Jurkowska, R. Z. & Jurkowski, T. P. Target specificity of mammalian DNA methylation and demethylation machinery. en. *Org. Biomol. Chem.* **16,** 1419–1435 (Feb. 2018).

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

9. Kohli, R. M. & Zhang, Y. TET enzymes, TDG and the dynamics of DNA demethylation. en. *Nature* **502,** 472–479 (Oct. 2013).

10. Wang, M., Zhang, K., Ngo, V., Liu, C., Fan, S., Whitaker, J. W., Chen, Y., Ai, R., Chen, Z. & Wang, J. Identification of DNA motifs that regulate DNA methylation. *BioRxiv,* 573352 (2019).

11. Whitaker, J. W., Chen, Z. & Wang, W. Predicting the human epigenome from DNA motifs. en. *Nat. Methods* **12,** 265–272, 7 p following 272 (Mar. 2015).

12. Wu, C., Yao, S., Li, X. & Chen Chujia and Hu, X. Genome-Wide Prediction of DNA Methylation Using DNA Composition and Sequence Complexity in Human. en. *Int. J. Mol. Sci.* **18** (Feb. 2017).

13. Das, R., Dimitrova, N., Xuan, Z., Rollins, R. A., Haghighi, F., Edwards, J. R., Ju, J., Bestor, T. H. & Zhang, M. Q. Computational prediction of methylation status in human genomic sequences. *Proceedings of the National Academy of Sciences of the United States of America* **103,** 10713–6. ISSN: 0027-8424 (July 2006).

14. Feng, P., Chen, W. & Lin, H. Prediction of CpG island methylation status by integrating DNA physicochemical properties. en. *Genomics* **104,** 229–233 (Oct. 2014).

15. Angermueller, C., Lee, H. J., Reik, W. & Stegle, O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. en. *Genome Biol.* **18,** 67 (Apr. 2017).

16. 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., Gingrich, J. A., Haghighi, F., Nutter, R. & Bestor, T. H. Chromatin and sequence features that define the fine and gross structure of genomic methylation patterns. en. *Genome Res.* **20,** 972–980 (July 2010).

17. Yamada, Y. & Satou, K. Prediction of genomic methylation status on CpG islands using DNA sequence features. *WSEAS Transactions on Biology and Biomedicine* **5,** 153–162 (2008).

18. Su, J., Shao, X., Liu, H., Liu, S., Wu, Q. & Zhang, Y. Genome-wide dynamic changes of DNA methylation of repetitive elements in human embryonic stem cells and fetal fibroblasts. en. *Genomics* **99,** 10–17 (Jan. 2012).

19. Wang, Y., Liu, T., Xu, D., Shi, H., Zhang, C., Mo, Y.-Y. & Wang, Z. Predicting DNA Methylation State of CpG Dinucleotide Using Genome Topological Features and Deep Networks. en. *Sci. Rep.* **6,** 19598 (Jan. 2016).

20. Wrzodek, C., Büchel, F., Hinselmann, G., Eichner, J., Mittag, F. & Zell, A. Linking the epigenome to the genome: correlation of different features to DNA methylation of CpG islands. en. *PLoS One* **7,** e35327 (Apr. 2012).

21. Zeng, H. & Gifford, D. K. Predicting the impact of non-coding variants on DNA methylation. en. *Nucleic Acids Res.* **45,** e99 (June 2017).

22. Lienert, F., Wirbelauer, C., Som, I., Dean, A., Mohn, F. & Schübeler, D. Identification of genetic elements that autonomously determine DNA methylation states. *Nat. Genet.* **43,** 1091–1097 (2011).

23. Stadler, M. B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., van Nimwegen, E., Wirbelauer, C., Oakeley, E. J., Gaidatzis, D., Tiwari, V. K. & Schübeler, D. DNA-binding

factors shape the mouse methylome at distal regulatory regions. *Nature* **480,** 490–5. ISSN: 1476-4687 (Dec. 2011).

24. Elango, N. & Yi, S. V. Functional relevance of CpG island length for regulation of gene expression. en. *Genetics* **187,** 1077–1083 (Apr. 2011).

25. Zhang, L., Gu, C., Yang, L., Tang, F. & Gao, Y. Q. The sequence preference of DNA methylation variation in mammalians. en. *PLoS One* **12,** e0186559 (Oct. 2017).

26. Fujiki, K., Shinoda, A., Kano, F., Sato, R., Shirahige, K. & Murata, M. PPARγ-induced PARylation promotes local DNA demethylation by production of 5-hydroxymethylcytosine. en. *Nat. Commun.* **4,** 2262 (2013).

27. Suzuki, T., Shimizu, Y., Furuhata, E., Maeda, S., Kishima, M., Nishimura, H., Enomoto Saaya and Hayashizaki, Y. & Suzuki, H. RUNX1 regulates site specificity of DNA demethylation by recruitment of DNA demethylation machineries in hematopoietic cells. en. *Blood Adv* **1,** 1699–1711 (Sept. 2017).

28. Suzuki, T., Maeda, S., Furuhata, E., Shimizu, Y., Nishimura, H., Kishima, M. & Suzuki, H. A screening system to identify transcription factors that induce binding site-directed DNA demethylation. *Epigenetics Chromatin* **10,** 60 (Dec. 2017).

29. Brandeis, M., Frank, D., Keshet, I., Siegfried, Z., Mendelsohn, M., Names, A., Temper, V., Razin, A. & Cedar, H. Spl elements protect a CpG island from de novo methylation. *Nature* **371,** 435–438 (Sept. 1994).

30. Macleod, D., Charlton, J., Mullins, J. & Bird, A. P. Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. *Genes Dev.* **8,** 2282–2292 (Oct. 1994).

31. Thomson, J. P., Skene, P. J., Selfridge, J., Clouaire, T., Guy, J., Webb, S., Kerr, A. R. W., Deaton, A., Andrews, R., James, K. D., Turner, D. J., Illingworth, R. & Bird, A. CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature* **464,** 1082–1086 (Apr. 2010).

32. Ko, M., An, J., Bandukwala, H. S., Chavez, L., Aijö, T., Pastor, W. A., Segal, M. F., Li, H., Koh, K. P., Lähdesmäki, H., Hogan, P. G., Aravind, L. & Rao, A. Modulation of TET2 expression and 5-methylcytosine oxidation by the CXXC domain protein IDAX. *Nature* **497,** 122–126 (May 2013).

33. Long, H. K., Blackledge, N. P. & Klose, R. J. ZF-CxxC domain-containing proteins, CpG islands and the chromatin connection. *Biochem. Soc. Trans.* **41,** 727–740 (June 2013).

34. Song, J., Rechkoblit, O., Bestor, T. H. & Patel, D. J. Structure of DNMT1-DNA complex reveals a role for autoinhibition in maintenance DNA methylation. *Science* **331,** 1036–1040 (Feb. 2011).

35. Zhang, D., Cheng, L., Badner, J. A., Chen, C., Chen, Q., Luo, W., Craig, D. W., Redman, M., Gershon, E. S. & Liu, C. Genetic control of individual differences in gene-specific methylation in human brain. *Am. J. Hum. Genet.* **86,** 411–419 (Mar. 2010).

36. Frauer, C., Rottach, A., Meilinger, D., Bultmann, S., Fellinger, K., Hasenöder, S., Wang, M., Qin, W., Söding, J., Spada, F. & Leonhardt, H. Different binding properties and function of CXXC zinc finger domains in Dnmt1 and Tet1. *PLoS One* **6,** e16627 (Feb. 2011).

37. Sato, N., Kondo, M. & Arai, K.-I. The orphan nuclear receptor GCNF recruits DNA methyltransferase for Oct-3/4 silencing. en. *Biochem. Biophys. Res. Commun.* **344,** 845–851 (June 2006).

38. Brenner, C., Deplus, R., Didelot, C., Loriot, A., Viré, E., De Smet, C., Gutierrez, A., Danovi, D., Bernard, D., Boon, T., Pelicci, P. G., Amati, B., Kouzarides, T., de Launoit, Y., Di Croce, L. & Fuks, F. Myc represses transcription through recruitment of DNA methyltransferase corepressor. en. *EMBO J.* **24,** 336–346 (Jan. 2005).

39. Velasco, G., Hubé, F., Rollin, J., Neuillet, D., Philippe, C., Bouzinba-Segard, H., Galvani, A., Viegas-Péquignot, E. & Francastel, C. Dnmt3b recruitment through E2F6 transcriptional repressor mediates germ-line gene silencing in murine somatic tissues. en. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 9281–9286 (May 2010).

40. Clark, J. & Smith, S. S. Secondary structure at a hot spot for DNA methylation in DNA from human breast cancers. *Cancer Genomics Proteomics* **5,** 241–251 (Sept. 2008).

41. Mao, S.-Q., Ghanbarian, A. T., Spiegel, J., Cuesta, S. M., Beraldi, D., Di Antonio, M., Marsico, G., Hänsel-Hertsch, R., Tannahill, D. & Balasubramanian, S. *DNA G-quadruplex structures mold the DNA methylome* 2018.

42. Mukherjee, A. K., Sharma, S. & Chowdhury, S. Non-duplex G-Quadruplex Structures Emerge as Mediators of Epigenetic Modifications. *Trends Genet.* **35,** 129–144 (Feb. 2019).

43. Mishra, S. K., Tawani, A., Mishra, A. & Kumar, A. G4IPDB: A database for G-quadruplex structure forming nucleic acid interacting proteins. *Sci. Rep.* **6,** 38144 (Dec. 2016).

44. Burge, S., Parkinson, G. N., Hazel, P., Todd, A. K. & Neidle, S. Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res.* **34,** 5402–5415 (Sept. 2006).

45. Di Salvo, M., Pinatel, E., Talà, A., Fondi, M., Peano, C. & Alifano, P. G4PromFinder: an algorithm for predicting transcription promoters in GC-rich bacterial genomes based on AT-rich elements and G-quadruplex motifs. *BMC Bioinformatics* **19,** 36 (Feb. 2018).

46. Suzuki, M., Yamada, T., Kihara-Negishi, F., Sakurai, T., Hara, E., Tenen, D. G., Hozumi, N. & Oikawa, T. Site-specific DNA methylation by a complex of PU.1 and Dnmt3a/b. en. *Oncogene* **25,** 2477–2488 (Apr. 2006).

47. De la Rica, L., de la Rica, L., Rodr\'\iguez-Ubreva, J., Garc\'\ia, M., Islam, A. B., Urquiza, J. M., Hernando, H., Christensen, J., Helin, K., Gómez-Vaquero, C. & Ballestar, E. *PU.1 target genes undergo Tet2-coupled demethylation and DNMT3b-mediated methylation in monocyte-to-osteoclast differentiation* 2013.

48. Schoenherr, C. J., Levorse, J. M. & Tilghman, S. M. CTCF maintains differential methylation at the Igf2/H19 locus. *Nat. Genet.* **33,** 66–69 (Jan. 2003).

49. Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S. & Aiden, E. L. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159,** 1665–1680. ISSN: 00928674 (Dec. 2014).

50. Splinter, E., Heath, H., Kooren, J., Palstra, R.-J., Klous, P., Grosveld, F., Galjart, N. & de Laat, W. CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes Dev.* **20,** 2349–2354 (Sept. 2006).

51. Weth, O., Paprotka, C., Günther, K., Schulte, A., Baierl, M., Leers, J., Galjart, N. & Renkawitz, R. CTCF induces histone variant incorporation, erases the H3K27me3 histone mark and opens chromatin. *Nucleic Acids Res.* **42,** 11941–11951 (Oct. 2014).

52. Nishiyama, A., Yamaguchi, L. & Nakanishi, M. Regulation of maintenance DNA methylation via histone ubiquitylation. *J. Biochem.* **159,** 9–15 (Jan. 2016).

53. Li, H., Rauch, T., Chen, Z.-X., Szabó, P. E., Riggs, A. D. & Pfeifer, G. P. The histone methyltransferase SETDB1 and the DNA methyltransferase DNMT3A interact directly and localize to promoters silenced in cancer cells. *J. Biol. Chem.* **281,** 19489–19500 (July 2006).

54. Schultz, D. C. *SETDB1: a novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins* 2002.

55. Frietze, S., O'Geen, H., Blahnik, K. R., Jin, V. X. & Farnham, P. J. ZNF274 recruits the histone methyltransferase SETDB1 to the 3' ends of ZNF genes. *PLoS One* **5,** e15082 (Dec. 2010).

56. Viré, E., Brenner, C., Deplus, R., Blanchon, L., Fraga, M., Didelot, C., Morey, L., Van Eynde, A., Bernard, D., Vanderwinden, J.-M., Bollen, M., Esteller, M., Di Croce, L., de Launoit, Y. & Fuks, F. The Polycomb group protein EZH2 directly controls DNA methylation. *Nature* **439,** 871–874 (Feb. 2006).

57. Baubec, T., Colombo, D. F., Wirbelauer Christiane and Schmidt, J., Burger, L., Krebs, A. R., Akalin, A. & Schübeler, D. Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. en. *Nature* **520,** 243–247 (Apr. 2015).

58. Clouaire, T., Webb, S., Skene, P., Illingworth, R., Kerr, A., Andrews, R., Lee, J.-H., Skalnik, D. & Bird, A. Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells. *Genes Dev.* **26,** 1714–1728 (Aug. 2012).

59. Law, J. A. & Jacobsen, S. E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* **11,** 204–220 (Mar. 2010).

60. Delaval, K., Govin, J., Cerqueira, F., Rousseaux, S., Khochbin, S. & Feil, R. Differential histone modifications mark mouse imprinting control regions during spermatogenesis. *EMBO J.* **26,** 720–729 (Feb. 2007).

61. Ciccone, D. N., Su, H., Hevi, S., Gay, F., Lei, H., Bajko, J., Xu, G., Li, E. & Chen, T. KDM1B is a histone H3K4 demethylase required to establish maternal genomic imprints. *Nature* **461,** 415–418 (Sept. 2009).

62. Rollins, R. A., Haghighi, F., Edwards, J. R., Das, R., Zhang, M. Q., Ju, J. & Bestor, T. H. Large-scale structure of genomic methylation patterns. *Genome Res.* **16,** 157–163 (Feb. 2006).

63. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. en. *Genes Dev.* **25,** 1010–1022 (May 2011).

64. Zheng, H., Wu, H., Li, J. & Jiang, S.-W. CpGIMethPred: computational model for predicting methylation status of CpG islands in human genome. *BMC medical genomics* **6 Suppl 1,** S13. ISSN: 1755-8794 (Jan. 2013).

65. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (Sept. 2012).

66. Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom Richard S and Eaton, M. L., Wu, Y.-C., Pfenning, A. R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R. A., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R. D., Lister, R., Hong Chibo and Gascard, P., Mungall, A. J., Moore Richard and Chuah, E., Tam, A., Canfield, T. K., Hansen, R. S., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh Kai-How and Feizi, S., Karlic, R., Kim, A.-R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Beaudet, A. E., Boyer Laurie A and De Jager, P. L., Farnham, P. J., Fisher Susan J and Haussler, D., Jones, S. J. M., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai, L.-H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker Joseph R and Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos,

J. A., Wang, T. & Kellis, M. Integrative analysis of 111 reference human epigenomes. en. *Nature* **518,** 317–330 (Feb. 2015).

67. Cancer Genome Atlas Research Network, Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C. & Stuart, J. M. The Cancer Genome Atlas Pan-Cancer analysis project. en. *Nat. Genet.* **45,** 1113–1120 (Oct. 2013).

68. Bujold, D., Morais, D. A. d. L., Gauthier, C., Côté, C., Caron, M., Kwan, T., Chen, K. C., Laperle, J., Markovits, A. N., Pastinen, T., Caron, B., Veilleux, A., Jacques, P.-É. & Bourque, G. The International Human Epigenome Consortium Data Portal. *Cell Syst* **3,** 496–499.e2 (Nov. 2016).

69. Ngo, V., Wang, M. & Wang, W. Finding de novo methylated DNA motifs. *Bioinformatics* (Feb. 2019).

70. Xuan Lin, Q. X., Sian, S., An, O., Thieffry, D., Jha, S. & Benoukraf, T. MethMotif: an integrative cell specific database of transcription factor binding motifs coupled with DNA methylation profiles. *Nucleic Acids Res.* **47,** D145–D154 (Jan. 2019).

71. Ngo, V., Chen, Z., Zhang, K., Whitaker, J. W., Wang, M. & Wang, W. Epigenomic analysis reveals DNA motifs regulating histone modifications in human and mouse. en. *Proc. Natl. Acad. Sci. U. S. A.* (Feb. 2019).

72. Dor, Y. & Cedar, H. Principles of DNA methylation and their implications for biology and medicine. *Lancet* **392,** 777–786 (Sept. 2018).

73. Fardi, M., Solali, S. & Farshdousti Hagh, M. Epigenetic mechanisms as a new approach in cancer treatment: An updated review. *Genes Dis* **5,** 304–311 (Dec. 2018).

74. Corcoran, R. B. & Chabner, B. A. Application of Cell-free DNA Analysis to Cancer Treatment. *N. Engl. J. Med.* **379,** 1754–1765 (Nov. 2018).

75. Wan, J. C. M., Massie, C., Garcia-Corbacho, J., Mouliere, F., Brenton, J. D., Caldas, C., Pacey, S., Baird, R. & Rosenfeld, N. Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nat. Rev. Cancer* **17,** 223–238 (Apr. 2017).

76. Hao, X., Luo, H., Krawczyk, M., Wei, W., Wang, W., Wang, J., Flagg, K., Hou, J., Zhang, H., Yi, S., Jafari, M., Lin, D., Chung, C., Caughey, B. A., Li, G., Dhar, D., Shi, W., Zheng, L., Hou, R., Zhu, J., Zhao, L., Fu, X., Zhang, E., Zhang, C., Zhu, J.-K., Karin, M. & Xu Rui-Hua and Zhang, K. DNA methylation markers for diagnosis and prognosis of common cancers. en. *Proc. Natl. Acad. Sci. U. S. A.* **114,** 7414–7419 (July 2017).

77. Underhill, H. R., Kitzman, J. O., Hellwig, S., Welker, N. C., Daza, R., Baker, D. N., Gligorich, K. M., Rostomily, R. C., Bronner, M. P. & Shendure, J. Fragment Length of Circulating Tumor DNA. *PLoS Genet.* **12,** e1006162 (July 2016).

78. Mouliere, F., Chandrananda, D., Piskorz, A. M., Moore, E. K., Morris, J., Ahlborn, L. B., Mair, R., Goranova, T., Marass, F., Heider, K., Wan, J. C. M., Supernat, A., Hudecova, I., Gounaris, I., Ros, S., Jimenez-Linan, M., Garcia-Corbacho, J., Patel, K., Østrup, O., Murphy, S., Eldridge, M. D., Gale, D., Stewart, G. D., Burge, J., Cooper, W. N., van der Heijden, M. S., Massie, C. E., Watts, C., Corrie, P., Pacey, S., Brindle, K. M., Baird, R. D., Mau-Sørensen, M., Parkinson, C. A., Smith, C. G., Brenton, J. D. & Rosenfeld, N. Enhanced detection of circulating tumor DNA by fragment size analysis. *Sci. Transl. Med.* **10** (Nov. 2018).

79. Lapin, M., Oltedal, S., Tjensvoll, K., Buhl, T., Smaaland, R., Garresori, H., Javle, M., Glenjen, N. I., Abelseth, B. K., Gilje, B. & Nordgård, O. Fragment size and level of cell-free DNA provide prognostic information in patients with advanced pancreatic cancer. *J. Transl. Med.* **16,** 300 (Nov. 2018).

80. Odegaard, J. I., Vincent, J. J., Mortimer, S., Vowles, J. V., Ulrich, B. C., Banks, K. C., Fairclough, S. R., Zill, O. A., Sikora, M., Mokhtari, R., Abdueva, D., Nagy, R. J., Lee, C. E., Kiedrowski, L. A., Paweletz, C. P., Eltoukhy, H., Lanman, R. B., Chudova, D. I. & Talasaz, A. Validation of a Plasma-Based Comprehensive Cancer Genotyping Assay Utilizing Orthogonal Tissue- and Plasma-Based Methodologies. *Clin. Cancer Res.* **24,** 3539–3549 (Aug. 2018).

81. Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M. & Shendure, J. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell* **164,** 57–68 (Jan. 2016).

82. Cohen, J. D., Li, L., Wang, Y., Thoburn, C., Afsari, B., Danilova, L., Douville, C., Javed, A. A., Wong, F., Mattox, A., Hruban, R. H., Wolfgang, C. L., Goggins, M. G., Dal Molin, M., Wang, T.-L., Roden, R., Klein, A. P., Ptak, J., Dobbyn, L., Schaefer, J., Silliman, N., Popoli, M., Vogelstein, J. T., Browne, J. D., Schoen, R. E., Brand, R. E., Tie, J., Gibbs, P., Wong, H.-L., Mansfield, A. S., Jen, J., Hanash, S. M., Falconi, M., Allen, P. J., Zhou, S., Bettegowda, C., Diaz, L., Tomasetti, C., Kinzler, K. W., Vogelstein, B., Lennon, A. M. & Papadopoulos, N. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* **3247,** eaar3247. ISSN: 0036-8075 (2018).

83. Gai, W. & Sun, K. Epigenetic Biomarkers in Cell-Free DNA and Applications in Liquid Biopsy. *Genes* **10** (Jan. 2019).

84. Ehrlich, M. DNA hypomethylation in cancer cells. en. *Epigenomics* **1,** 239–259 (Dec. 2009).

85. Müller, H. M., Widschwendter, A., Fiegl, H., Ivarsson, L., Goebel, G., Perkmann, E., Marth, C. & Widschwendter, M. DNA methylation in serum of breast cancer patients: an independent prognostic marker. *Cancer Res.* **63,** 7641–7645 (Nov. 2003).

86. Fiegl, H., Millinger, S., Mueller-Holzner, E., Marth, C., Ensinger, C., Berger, A., Klocker, H., Goebel, G. & Widschwendter, M. Circulating tumor-specific DNA: a marker for monitoring efficacy of adjuvant therapy in cancer patients. *Cancer Res.* **65,** 1141–1145 (Feb. 2005).

87. Fackler, M. J., Lopez Bujanda, Z., Umbricht, C., Teo, W. W., Cho, S., Zhang, Z., Visvanathan, K., Jeter, S., Argani, P., Wang, C., Lyman, J. P., de Brot, M., Ingle, J. N., Boughey, J., McGuire, K., King, T. A., Carey, L. A., Cope, L., Wolff, A. C. & Sukumar, S. Novel methylated biomarkers and a robust assay to detect circulating tumor DNA in metastatic breast cancer. *Cancer Res.* **74,** 2160–2170 (Apr. 2014).

88. Henriksen, S. D., Madsen, P. H., Larsen, A. C., Johansen, M. B., Pedersen, I. S., Krarup, H. & Thorlacius-Ussing, O. *Promoter hypermethylation in plasma-derived cell-free DNA as a prognostic marker for pancreatic adenocarcinoma staging* 2017.

89. Widschwendter, M., Evans, I., Jones, A., Ghazali, S., Reisel, D., Ryan, A., Gentry-Maharaj, A., Zikan, M., Cibula, D., Eichner, J., Alunni-Fabbroni, M., Koch, J., Janni, W. J., Paprotka, T., Wittenberger, T., Menon, U., Wahl, B., Rack, B. & Lempiäinen, H. Methylation patterns in serum DNA for early identification of disseminated breast cancer. *Genome Med.* **9,** 115 (Dec. 2017).

90. Zhao, F., Olkhov-Mitsel, E., Kamdar, S., Jeyapala, R., Garcia, J., Hurst, R., Hanna, M. Y., Mills, R., Tuzova, A. V., O'Reilly, E., Kelly, S., Cooper, C., Movember Urine Biomarker Consortium, Brewer, D., Perry, A. S., Clark, J., Fleshner, N. & Bapat, B. A urine-based DNA methylation assay, ProCUrE, to identify clinically significant prostate cancer. *Clin. Epigenetics* **10,** 147 (Nov. 2018).

91. Brikun, I., Nusskern, D., Decatus, A., Harvey, E., Li, L. & Freije, D. A panel of DNA methylation markers for the detection of prostate cancer from FV and DRE urine DNA. *Clin. Epigenetics* **10,** 91 (July 2018).

92. Han, Y. D., Oh, T. J., Chung, T.-H., Jang, H. W., Kim, Y. N., An, S. & Kim, N. K. Early detection of colorectal cancer based on presence of methylated syndecan-2 (SDC2) in stool DNA. *Clin. Epigenetics* **11,** 51 (Mar. 2019).

93. Guo, S., Diep, D., Plongthongkum, N., Fung, H.-L., Zhang, K. & Zhang, K. Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. en. *Nat. Genet.* **49,** 635–642 (Apr. 2017).

94. Moss, J., Magenheim, J., Neiman, D., Zemmour, H., Loyfer, N., Korach, A., Samet, Y., Maoz, M., Druid, H., Arner, P., Fu, K.-Y., Kiss, E., Spalding, K. L., Landesberg, G., Zick, A., Grinshpun, A., Shapiro, A. M. J., Grompe, M., Wittenberg, A. D., Glaser, B., Shemer, R., Kaplan, T. & Dor, Y. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nat. Commun.* **9,** 5068 (Nov. 2018).

95. Shen, S. Y., Singhania, R., Fehringer, G., Chakravarthy, A., Roehrl, M. H. A., Chadwick, D., Zuzarte, P. C., Borgida, A., Wang, T. T., Li, T., Kis, O., Zhao, Z., Spreafico, A., Medina, T. d. S., Wang, Y., Roulois, D., Ettayebi, I., Chen, Z., Chow, S., Murphy, T., Arruda, A., O'Kane, G. M., Liu, J., Mansour, M., McPherson, J. D., O'Brien, C., Leighl, N., Bedard, P. L., Fleshner, N., Liu, G., Minden, M. D., Gallinger, S., Goldenberg, A., Pugh, T. J., Hoffman, M. M., Bratman, S. V., Hung, R. J. & De Carvalho, D. D. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature* **563,** 579–583 (Nov. 2018).

96. Rapkins, R. W., Wang, F., Nguyen, H. N., Cloughesy, T. F., Lai, A., Ha, W., Nowak, A. K., Hitchins, M. P. & McDonald, K. L. The MGMT promoter SNP rs16906252 is a risk factor for MGMT methylation in glioblastoma and is predictive of response to temozolomide. en. *Neuro. Oncol.* **17,** 1589–1598 (Dec. 2015).

97. De Toro-Martin, J., Guenard, F., Tchernof, A., Deshaies, Y., Perusse, L., Biron, S., Lescelleur, O., Biertho, L., Marceau, S. & Vohl, M.-C. A CpG-SNP located within the ARPC3 gene promoter is associated with hypertriglyceridemia in severely obese patients. *Ann. Nutr. Metab.* **68,** 203–212 (2016).

98. Shilpi, A., Bi, Y., Jung, S., Patra, S. K. & Davuluri, R. V. Identification of Genetic and Epigenetic Variants Associated with Breast Cancer Prognosis by Integrative Bioinformatics Analysis. *Cancer Inform.* **16,** 1–13 (Jan. 2017).

99. Fan, H., Liu, D., Qiu, X., Qiao, F., Wu, Q., Su, X., Zhang, F., Song, Y., Zhao, Z. & Xie, W. A functional polymorphism in the DNA methyltransferase-3A promoter modifies the susceptibility in gastric cancer but not in esophageal carcinoma. *BMC Med.* **8,** 12 (Feb. 2010).

100. Rakyan, V. K., Hildmann, T., Novik, K. L., Lewin, J., Tost, J., Cox, A. V., Andrews, T. D., Howe, K. L., Otto, T., Olek, A., Fischer, J., Gut, I. G., Berlin, K. & Beck, S. DNA methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project. *PLoS Biol.* **2,** e405 (Dec. 2004).

101. Kerkel, K., Spadola, A., Yuan, E., Kosek, J., Jiang, L., Hod, E., Li, K., Murty, V. V., Schupf, N., Vilain, E., Morris, M., Haghighi, F. & Tycko, B. Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. *Nat. Genet.* **40,** 904–908 (July 2008).

102. Shoemaker, R., Deng, J., Wang, W. & Zhang, K. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome Res.* **20,** 883–889 (July 2010).

103. Amin, V., Harris, R. A., Onuchic, V., Jackson, A. R., Charnecki, T., Paithankar, S., Lakshmi Subramanian, S., Riehle, K., Coarfa, C. & Milosavljevic, A. Epigenomic footprints across

111 reference epigenomes reveal tissue-specific epigenetic regulation of lincRNAs. en. *Nat. Commun.* **6,** 6370 (Feb. 2015).

104. Kulakovskiy, I. V., Vorontsov, I. E., Yevshin, I. S., Sharipov, R. N., Fedorova, A. D., Rumynskiy Eugene I and Medvedeva, Y. A., Magana-Mora, A., Bajic, V. B., Papatsenko, D. A., Kolpakov, F. A. & Makeev, V. J. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. en. *Nucleic Acids Res.* **46,** D252–D259 (Jan. 2018).

105. Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W. & Noble, W. S. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research* **37,** W202–W208. ISSN: 0305-1048 (July 2009).

106. SOKAL & R, R. A statistical method for evaluating systematic relationship. *University of Kansas Science Bulletin* **28,** 1409–1438 (1958).

107. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. en. *Bioinformatics* **27,** 1017–1018 (Apr. 2011).

108. Deplus, R., Delatte, B., Schwinn, M. K., Defrance, M., Méndez, J., Murphy, N., Dawson, M. A., Volkmar, M., Putmans, P., Calonne, E., Shih, A. H., Levine, R. L., Bernard, O., Mercher, T., Solary, E., Urh, M., Daniels, D. L. & Fuks, F. TET2 and TET3 regulate GlcNAcylation and H3K4 methylation through OGT and SET1/COMPASS. en. *EMBO J.* **32,** 645–655 (Mar. 2013).

109. Jin, B., Ernst, J., Tiedemann, R. L., Xu, H., Sureshchandra, S., Kellis, M., Dalton, S., Liu, C., Choi, J.-H. & Robertson, K. D. Linking DNA methyltransferases to epigenetic marks and nucleosome structure genome-wide in human tumor cells. en. *Cell Rep.* **2,** 1411–1424 (Nov. 2012).

110. Verma, N., Pan, H., Doré, L. C., Shukla, A., Li, Q. V., Pelham-Webb, B., Teijeiro, V., González, F., Krivtsov, A., Chang, C.-J., Papapetrou, E. P., He, C., Elemento, O. & Huangfu, D. TET proteins safeguard bivalent promoters from de novo methylation in human embryonic stem cells. en. *Nat. Genet.* **50,** 83–95 (Jan. 2018).

111. Kemp, C. J., Moore, J. M., Moser, R., Bernard, B., Teater, M., Smith, L. E., Rabaia, N. A., Gurley, K. E., Guinney, J., Busch, S. E., Shaknovich, R., Lobanenkov, V. V., Liggitt, D., Shmulevich, I., Melnick, A. & Filippova, G. N. CTCF haploinsufficiency destabilizes DNA methylation and predisposes to cancer. en. *Cell Rep.* **7,** 1020–1029 (May 2014).

112. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. en. *Bioinformatics* **28,** 1353–1358 (May 2012).

113. Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., Steward, C., Harte,

R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N., Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J. M., Ezkurdia, I., van Baren, J., Brent, M., Haussler, D., Kellis, M., Valencia, A., Reymond, A., Gerstein, M., Guigó, R. & Hubbard, T. J. GENCODE: the reference human genome annotation for The ENCODE Project. en. *Genome Res.* **22,** 1760–1774 (Sept. 2012).

114. Gaunt, T. R., Shihab, H. A., Hemani, G., Min, J. L., Woodward, G., Lyttleton, O., Zheng Jie and Duggirala, A., McArdle, W. L., Ho, K., Ring, S. M., Evans, D. M., Davey Smith, G. & Relton, C. L. Systematic identification of genetic influences on methylation across the human life course. en. *Genome Biol.* **17,** 61 (Mar. 2016).

115. Gutierrez-Arcelus, M., Lappalainen, T., Montgomery, S. B., Buil, A., Ongen, H., Yurovsky Alisa and Bryois, J., Giger, T., Romano, L., Planchon, A., Falconnet, E., Bielser Deborah and Gagnebin, M., Padioleau, I., Borel, C., Letourneau, A., Makrythanasis, P., Guipponi, M., Gehrig, C. & Antonarakis Stylianos E and Dermitzakis, E. T. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. en. *Elife* **2,** e00523 (June 2013).

116. Hannon, E., Spiers, H., Viana, J., Pidsley, R., Burrage, J., Murphy, T. M., Troakes Claire and Turecki, G., O'Donovan, M. C., Schalkwyk, L. C., Bray, N. J. & Mill, J. Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. en. *Nat. Neurosci.* **19,** 48–54 (Jan. 2016).

117. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **29,** 1189–1232 (Oct. 2001).

118. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, É. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12,** 2825–2830 (Feb. 2011).

119. Olson, R. S., La Cava, W., Mustahsan Zairah and Varik, A. & Moore, J. H. Data-driven Advice for Applying Machine Learning to Bioinformatics Problems. *arXiv[q-bio].* arXiv: `1708.05070` (Aug. 2017).

120. Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Tamborero, D., Schroeder, M. P., Jene-Sanz, A., Santos, A. & Lopez-Bigas, N. IntOGen-mutations identifies cancer drivers across tumor types. en. *Nat. Methods* **10,** 1081–1082 (Nov. 2013).

121. Therneau, T. M. & Lumley, T. Package 'survival'. *R Top Doc* **128** (2015).

122. 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., Lin, S., Lin, Y., Jung, I., Schmitt, A. D., Selvaraj, S., Ren, B., Sejnowski, T. J., Wang, W. & Ecker, J. R. Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* **523,** 212–216. ISSN: 0028-0836 (June 2015).

123. Jeong, M., Sun, D., Luo, M., Huang, Y., Challen, G. A., Rodriguez, B., Zhang, X., Chavez, L., Wang, H., Hannah, R., Kim Sang-Bae and Yang, L., Ko, M., Chen, R., Göttgens, B., Lee, J.-S., Gunaratne, P., Godley, L. A., Darlington, G. J., Rao, A., Li, W. & Goodell, M. A. Large conserved domains of low DNA methylation maintained by Dnmt3a. en. *Nat. Genet.* **46,** 17–23 (Jan. 2014).

124. Witte, T., Plass, C. & Gerhauser, C. Pan-cancer patterns of DNA methylation. en. *Genome Med.* **6,** 66 (Aug. 2014).

125. Hovestadt, V., Jones, D. T. W., Picelli, S., Wang, W., Kool, M., Northcott, P. A., Sultan Marc and Stachurski, K., Ryzhova, M., Warnatz, H.-J., Ralser, M., Brun, S., Bunt Jens and Jäger, N., Kleinheinz, K., Erkek Serap and Weber, U. D., Bartholomae, C. C., von Kalle, C., Lawerenz, C., Eils, J., Koster Jan and Versteeg, R., Milde, T., Witt, O., Schmidt, S., Wolf, S., Pietsch, T., Rutkowski, S., Scheurlen, W., Taylor, M. D., Brors, B., Felsberg, J., Reifenberger, G., Borkhardt, A., Lehrach, H., Wechsler-Reya Robert J and Eils, R., Yaspo, M.-L., Landgraf, P., Korshunov, A., Zapatka, M., Radlwimmer, B., Pfister, S. M. & Lichter, P. Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing. en. *Nature* **510,** 537–541 (June 2014).

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

127. Gu, J., Stevens, M., Xing, X., Li, D., Zhang, B., Payton, J. E., Oltz Eugene M and Jarvis, J. N., Jiang, K., Cicero, T., Costello, J. F. & Wang, T. Mapping of Variable DNA Methylation Across Multiple Cell Types Defines a Dynamic Regulatory Landscape of the Human Genome. en. *G3* **6,** 973–986 (Apr. 2016).

128. McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B. & Wenger Aaron M and Bejerano, G. GREAT improves functional interpretation of cis-regulatory regions. en. *Nat. Biotechnol.* **28,** 495–501 (May 2010).

129. Maunakea, A. K., Chepelev, I., Cui, K. & Zhao, K. Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. en. *Cell Res.* **23,** 1256–1269 (Nov. 2013).

130. Swami, M. Epigenetics: Demethylation links cell fate and cancer. en. *Nat. Rev. Cancer* **10,** 740 (Nov. 2010).

131. Bagci, H. & Fisher, A. G. DNA demethylation in pluripotency and reprogramming: the role of tet proteins and cell division. en. *Cell Stem Cell* **13,** 265–269 (Sept. 2013).

132. Kumar, N., Hoque, M. A. & Sugimoto, M. Robust volcano plot: identification of differential metabolites in the presence of outliers. en. *BMC Bioinformatics* **19,** 128 (Apr. 2018).

133. Giambra, V., Volpi, S., Emelyanov Alexander V and Pflugh, D., Bothwell, A. L. M., Norio, P., Fan, Y., Ju, Z., Skoultchi, A. I., Hardy, R. R., Frezza, D. & Birshtein, B. K. Pax5 and linker histone H1 coordinate DNA methylation and histone modifications in the 3' regulatory region of the immunoglobulin heavy chain locus. en. *Mol. Cell. Biol.* **28,** 6123–6133 (Oct. 2008).

134. Maag, J. L. V., Kaczorowski, D. C., Panja Debabrata and Peters, T. J., Bramham, C. R., Wibrand, K. & Dinger, M. E. Widespread promoter methylation of synaptic plasticity genes in long-term potentiation in the adult brain in vivo. en. *BMC Genomics* **18,** 250 (Mar. 2017).

135. Yevshin, I., Sharipov, R., Kolmykov, S., Kondrakhin, Y. & Kolpakov, F. GTRD: a database on gene transcription regulation-2019 update. *Nucleic Acids Res.* **47,** D100–D105 (Jan. 2019).

136. Cuozzo, C., Porcellini, A., Angrisano Tiziana and Morano, A., Lee, B., Di Pardo, A., Messina, S., Iuliano, R., Fusco, A., Santillo, M. R., Muller, M. T., Chiariotti, L., Gottesman, M. E. & Avvedimento, E. V. DNA damage, homology-directed repair, and DNA methylation. en. *PLoS Genet.* **3,** e110 (July 2007).

137. De la Rica, L., Deniz, Ö., Cheng, K. C. L., Todd, C. D., Cruz, C., Houseley, J. & Branco, M. R. TET-dependent regulation of retrotransposable elements in mouse embryonic stem cells. en. *Genome Biol.* **17,** 234 (Nov. 2016).

138. Luu, P. L., Scholer, H. R. & Arauzo-Bravo, M. J. Disclosing the crosstalk among DNA methylation, transcription factors, and histone marks in human pluripotent cells through discovery of DNA methylation motifs. *Genome Res* **23,** 2013–2029 (2013).

139. Jones - Nature Reviews Genetics, P. A. & 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *nature.com* (2012).

140. Network, C. G. A. R. *et al.* Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45,** 1113–1120 (2013).

141. Wiehle, L., Raddatz, G., Musch, T., Dawlaty, M. M., Jaenisch, R., Lyko, F. & Breiling, A. Tet1 and Tet2 Protect DNA Methylation Canyons against Hypermethylation. en. *Mol. Cell. Biol.* **36,** 452–461 (Feb. 2016).

142. Okano, M., Bell, D. W., Haber, D. A. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. en. *Cell* **99,** 247–257 (Oct. 1999).

143. Morselli, M., Pastor, W. A., Montanini, B., Nee, K., Ferrari, R., Fu, K., Bonora Giancarlo and Rubbi, L., Clark, A. T., Ottonello Simone and Jacobsen, S. E. & Pellegrini, M. In vivo

targeting of de novo DNA methylation by histone modifications in yeast and mouse. en. *Elife* **4,** e06205 (Apr. 2015).

144. Duymich, C. E., Charlet, J., Yang Xiaojing and Jones, P. A. & Liang, G. DNMT3B isoforms without catalytic activity stimulate gene body methylation as accessory proteins in somatic cells. en. *Nat. Commun.* **7,** 11453 (Apr. 2016).

145. Challen, G. A., Sun, D., Mayle, A., Jeong, M., Luo, M., Rodriguez, B., Mallaney, C., Celik, H., Yang, L., Xia, Z., Cullen, S., Berg, J., Zheng, Y., Darlington, G. J., Li, W. & Goodell, M. A. Dnmt3a and Dnmt3b have overlapping and distinct functions in hematopoietic stem cells. en. *Cell Stem Cell* **15,** 350–364 (Sept. 2014).

146. Dyrvig, M., Qvist, P., Lichota, J., Larsen Knud and Nyegaard, M., Børglum, A. D. & Christensen, J. H. DNA Methylation Analysis of BRD1 Promoter Regions and the Schizophrenia rs138880 Risk Allele. en. *PLoS One* **12,** e0170121 (Jan. 2017).

147. Chuang, T.-J., Chen, F.-C. & Chen, Y.-Z. Position-dependent correlations between DNA methylation and the evolutionary rates of mammalian coding exons. en. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 15841–15846 (Sept. 2012).

148. De Leon, M., Cardenas, H., Vieth, E., Emerson, R., Segar, M., Liu, Y., Nephew, K. & Matei, D. Transmembrane protein 88 (TMEM88) promoter hypomethylation is associated with platinum resistance in ovarian cancer. en. *Gynecol. Oncol.* **142,** 539–547 (Sept. 2016).

149. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. en. *PLoS One* **10,** e0118432 (Mar. 2015).

150. Koukoura, O., Spandidos, D. A., Daponte, A. & Sifakis, S. DNA methylation profiles in ovarian cancer: implication in diagnosis and therapy (Review). en. *Mol. Med. Rep.* **10,** 3–9 (July 2014).

151. Ellis, R. J., Wang, Y., Stevenson, H. S., Boufraqech, M., Patel, D., Nilubol, N., Davis, S., Edelman, D. C., Merino, M. J., He Mei and Zhang, L., Meltzer, P. S. & Kebebew, E. Genome-wide methylation patterns in papillary thyroid cancer are distinct based on histological subtype and tumor genotype. en. *J. Clin. Endocrinol. Metab.* **99,** E329–37 (Feb. 2014).

152. Kroeger, H., Jelinek, J., Estécio Marcos R H and He, R., Kondo, K., Chung, W., Zhang Li and Shen, L., Kantarjian, H. M., Bueso-Ramos, C. E. & Issa, J.-P. J. Aberrant CpG island methylation in acute myeloid leukemia is accentuated at relapse. en. *Blood* **112,** 1366–1373 (Aug. 2008).

153. De Cubas, A. A., Korpershoek, E., Inglada-Pérez, L., Letouzé, E., Currás-Freixes, M., Fernández, A. F., Comino-Méndez, I., Schiavi, F., Mancikova, V., Eisenhofer, G., Mannelli Massimo and Opocher, G., Timmers, H., Beuschlein Felix and de Krijger, R., Cascon, A.,

Rodr\'\iguez-Antona, C., Fraga, M. F., Favier, J., Gimenez-Roqueplo, A.-P. & Robledo, M. DNA Methylation Profiling in Pheochromocytoma and Paraganglioma Reveals Diagnostic and Prognostic Markers. en. *Clin. Cancer Res.* **21,** 3020–3030 (July 2015).

154. Rondelet, G., Dal Maso, T., Willems, L. & Wouters, J. Structural basis for recognition of histone H3K36me3 nucleosome by human de novo DNA methyltransferases 3A and 3B. en. *J. Struct. Biol.* **194,** 357–367 (June 2016).

155. Rinaldi, L., Datta, D., Serrat, J., Morey, L., Solanas, G., Avgustinova, A., Blanco, E., Pons, J. I., Matallanas, D., Von Kriegsheim, A., Di Croce, L. & Benitah, S. A. Dnmt3a and Dnmt3b Associate with Enhancers to Regulate Human Epidermal Stem Cell Homeostasis. en. *Cell Stem Cell* **19,** 491–501 (Oct. 2016).

156. Liu, X., Wang, C., Liu, W., Li Jingyi and Li, C., Kou, X., Chen, J., Zhao Yanhong and Gao, H., Wang, H., Zhang, Y., Gao, Y. & Gao, S. Distinct features of H3K4me3 and H3K27me3 chromatin domains in pre-implantation embryos. en. *Nature* **537,** 558–562 (Sept. 2016).

157. Mangan, S. & Alon, U. Structure and function of the feed-forward loop network motif. en. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 11980–11985 (Oct. 2003).

158. Gu, T., Lin, X., Cullen, S. M., Luo, M., Jeong, M., Estecio, M., Shen, J., Hardikar, S., Sun, D., Su, J., Rux, D., Guzman, A., Lee, M., Qi, L. S., Chen, J.-J., Kyba, M., Huang, Y., Chen, T., Li, W. & Goodell, M. A. DNMT3A and TET1 cooperate to regulate promoter epigenetic landscapes in mouse embryonic stem cells. en. *Genome Biol.* **19,** 88 (July 2018).

159. Biergans, S. D., Giovanni Galizia, C., Reinhard, J. & Claudianos, C. Dnmts and Tet target memory-associated genes after appetitive olfactory training in honey bees. en. *Sci. Rep.* **5,** 16223 (Nov. 2015).

160. Jacob, S. T. & Motiwala, T. Epigenetic regulation of protein tyrosine phosphatases: potential molecular targets for cancer therapy. en. *Cancer Gene Ther.* **12,** 665–672 (Aug. 2005).

161. Aguiar, R. C., Yakushijin, Y., Kharbanda, S., Tiwari, S., Freeman, G. J. & Shipp, M. A. PTPROt: an alternatively spliced and developmentally regulated B-lymphoid phosphatase that promotes G0/G1 arrest. en. *Blood* **94,** 2403–2413 (Oct. 1999).

162. Wang, W., Cherry, J. M., Nochomovitz, Y., Jolly, E., Botstein, D. & Li, H. Inference of combinatorial regulation in yeast transcriptional networks: a case study of sporulation. en. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 1998–2003 (Feb. 2005).

163. Miller, J. A. & Widom, J. Collaborative competition mechanism for gene activation in vivo. en. *Mol. Cell. Biol.* **23,** 1623–1632 (Mar. 2003).

164. Darieva, Z., Clancy, A., Bulmer, R., Williams, E., Pic-Taylor, A., Morgan, B. A. & Sharrocks, A. D. A competitive transcription factor binding mechanism determines the timing of late cell cycle-dependent gene expression. en. *Mol. Cell* **38,** 29–40 (Apr. 2010).

165. Zhu, H., Wang, G. & Qian, J. Transcription factors as readers and effectors of DNA methylation. *Nature reviews. Genetics* **17.** ISSN: 1471-0064 (2016).

166. Xin, B. & Rohs, R. Relationship between histone modifications and transcription factor binding is protein family specific. *Genome Res.* (Jan. 2018).

167. Zhu, F., Farnung, L., Kaasinen, E., Sahu, B., Yin, Y., Wei, B., Dodonova, S. O., Nitta, K. R., Morgunova, E., Taipale, M., Cramer, P. & Taipale, J. The interaction landscape between transcription factors and the nucleosome. *Nature* **562,** 76–81 (Oct. 2018).

168. Schneider, T. D. & Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucleic acids research* **18,** 6097–100. ISSN: 0305-1048 (Oct. 1990).

169. Fletez-Brant, C., Lee, D., McCallion, A. S. & Beer, M. A. kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets. *Nucleic Acids Res.* **41,** W544–56 (July 2013).

170. Ghandi, M., Lee, D., Mohammad-Noori, M. & Beer, M. A. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput. Biol.* **10,** e1003711 (July 2014).

171. Guo, Y., Tian, K., Zeng, H., Guo, X. & Gifford, D. K. A novel k-mer set memory (KSM) motif representation improves regulatory variant prediction. *Genome Res.* **28,** 891–900 (June 2018).

172. Zeng, H., Hashimoto, T., Kang, D. D. & Gifford, D. K. GERV: a statistical method for generative evaluation of regulatory variants for transcription factor binding. *Bioinformatics* **32,** 490–496 (Feb. 2016).

173. Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. & Haussler, D. The human genome browser at UCSC. *Genome Res.* **12,** 996–1006 (June 2002).

174. Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G. & Mesirov, J. P. Integrative genomics viewer. *Nature Biotechnology* **29,** 24–26. ISSN: 1087-0156 (Jan. 2011).

175. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H. & Glass, C. K. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38,** 576–589 (May 2010).

176. Cavener, D. R. Comparison of the consensus sequence flanking translational start sites in Drosophila and vertebrates. **15** (1987).

177. Lin, J. Divergence Measures Based on the Shannon Entropy. *IEEE TRANSACTIONS ON INFORMATION THEORY* **37** (1991).

178. Lele, S. Euclidean Distance Matrix Analysis (EDMA): Estimation of mean form and mean form difference. *Math. Geol.* **25,** 573–602 (July 1993).

179. Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A. E. & Wingender, E. TRANSFAC\textregistered and its module TRANSCompel\textregistered: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34,** D108–D110 (Jan. 2006).

180. Portales-Casamar, E., Thongjuea, S., Kwon, A. T., Arenillas, D., Zhao, X., Valen, E., Yusuf, D., Lenhard, B., Wasserman, W. W. & Sandelin, A. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **38,** D105–10 (Jan. 2010).

181. Robasky, K. & Bulyk, M. L. UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein–DNA interactions. *Nucleic Acids Res.* **39,** D124–D128 (Jan. 2011).

182. Xie, Z., Hu, S., Blackshaw, S., Zhu, H. & Qian, J. hPDI: a database of experimental human protein–DNA interactions. *Bioinformatics* **26,** 287–289 (Jan. 2010).

183. Davis, J. & Goadrich, M. *The Relationship Between Precision-Recall and ROC Curves* in *Proceedings of the 23rd International Conference on Machine Learning* (ACM, New York, NY, USA, 2006), 233–240.

184. Johnson, A. D. An extended IUPAC nomenclature code for polymorphic nucleic acids. *Bioinformatics (Oxford, England)* **26,** 1386–9. ISSN: 1367-4811 (May 2010).

185. Bailey, T. L., Elkan, C., *et al.* Fitting a mixture model by expectation maximization to discover motifs in bipolymers (1994).

186. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays. *Genomics* **106,** 159–164. ISSN: 0888-7543 (2015).

187. Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, Ł. M., Rath, M. & Stark, A. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339,** 1074–1077. ISSN: 0036-8075 (2013).

188. Santiago-Algarra, D., Dao, L. T. M., Pradel, L., España, A. & Spicuglia, S. Recent advances in high-throughput approaches to dissect enhancer function. *F1000Res.* **6,** 939. ISSN: 2046-1402 (2017).

189. Adli, M. The CRISPR tool kit for genome editing and beyond. *Nat. Commun.* **9,** 1911. ISSN: 2041-1723 (2018).

190. Diao, Y., Li, B., Meng, Z., Jung, I., Lee, A. Y., Dixon, J., Maliskova, L., Guan, K.-L., Shen, Y. & Ren, B. A new class of temporarily phenotypic enhancers identified by CRISPR/Cas9-mediated genetic screening. *Genome Res.* **26,** 397–405. ISSN: 1088-9051 (2016).

191. Diao, Y., Fang, R., Li, B., Meng, Z., Yu, J., Qiu, Y., Lin, K. C., Huang, H., Liu, T., Marina, R. J., Jung, I., Shen, Y., Guan, K.-L. & Ren, B. A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nat. Methods* **14,** 629–635. ISSN: 1548-7091 (2017).

192. Li, W., Xu, H., Xiao, T., Cong, L., Love, M. I., Zhang, F., Irizarry, R. A., Liu, J. S., Brown, M. & Liu, X. S. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol.* **15,** 554. ISSN: 1465-6906 (2014).

193. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760. ISSN: 1367-4803 (2009).

194. Langmead, B. Aligning short sequencing reads with Bowtie. *Curr. Protoc. Bioinformatics* **Chapter 11,** Unit 11.7. ISSN: 1934-3396 (2010).

195. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359. ISSN: 1548-7091 (2012).

196. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842. ISSN: 1367-4803 (2010).

197. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140. ISSN: 1367-4803 (2010).

198. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11,** R25. ISSN: 1465-6906 (2010).

199. Kolde, R., Laur, S., Adler, P. & Vilo, J. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* **28,** 573–580. ISSN: 1367-4803 (2012).

200. Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W. & Liu, X. S. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137. ISSN: 1465-6906 (2008).

201. Liaw, A. & Wiener, M. Classification and regression based on a forest of trees using random inputs. *R Package LB - iwXO.*

202. Hariprakash, J. M. & Ferrari, F. Computational Biology Solutions to Identify Enhancers-target Gene Pairs. *Comput. Struct. Biotechnol. J.* **17,** 821–831. ISSN: 2001-0370 (2019).

203. Zhu, Y., Chen, Z., Zhang, K., Wang, M., Medovoy, D., Whitaker, J., Ding, B., Li, N., Zheng, L. & Wang, W. Constructing 3D interaction maps from 1D epigenomes. *Nature Communications* **7.** ISSN: 20411723 (2016).

204. Muers, M. Technology: Getting Moore from DNA sequencing. *Nat. Rev. Genet.* **12,** 586 (Aug. 2011).

205. Sicklick, J. K., Kato, S., Okamura, R., Schwaederle, M., Hahn, M. E., Williams, C. B., De, P., Krie, A., Piccioni, D. E., Miller, V. A., Ross, J. S., Benson, A., Webster, J., Stephens, P. J., Lee, J. J., Fanta, P. T., Lippman, S. M., Leyland-Jones, B. & Kurzrock, R. Molecular profiling of cancer patients enables personalized combination therapy: the I-PREDICT study. *Nat. Med.* **25,** 744–750 (May 2019).

206. Fiala, C. & Diamandis, E. P. Utility of circulating tumor DNA in cancer diagnostics with emphasis on early detection. *BMC Med.* **16,** 166 (Oct. 2018).

207. Koch, A., Joosten, S. C., Feng, Z., de Ruijter, T. C., Draht, M. X., Melotte, V., Smits, K. M., Veeck, J., Herman, J. G., Van Neste, L., Van Criekinge, W., De Meyer, T. & van Engeland, M. Analysis of DNA methylation in cancer: location revisited. *Nat. Rev. Clin. Oncol.* **15,** 459–466 (July 2018).

208. Zhang, L., Liang, Y., Li, S., Zeng, F., Meng, Y., Chen, Z., Liu, S., Tao, Y. & Yu, F. The interplay of circulating tumor DNA and chromatin modification, therapeutic resistance, and metastasis. *Mol. Cancer* **18,** 36 (Mar. 2019).

209. De, S. & Michor, F. DNA secondary structures and epigenetic determinants of cancer genome evolution. *Nat. Struct. Mol. Biol.* **18,** 950–955 (July 2011).