UNIVERSITY OF CALIFORNIA, SAN DIEGO

Beyond DNA sequence: Exploring and Exploiting Mammalian DNA Methylation

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

in

Bioinformatics and Systems Biology

by

Yupeng He

Committee in charge:

Professor Joseph R. Ecker, Chair
Professor Wei Wang, Co-Chair
Professor Vineet Bafna
Professor Bing Ren
Professor Kun Zhang

2017

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

Co-Chair

Chair

University of California, San Diego

2017

DEDICATION

To my family, especially my mother and father who have been supporting me to pursue a career to exploring the beauty and elegence of biology.

EPIGRAPH

Essentially, all models are wrong, but some are useful.

—George E. P. Box

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF SUPPLEMENTARY FILES

- He_chapter4_Supplemental_Table_3.xlsx
- He_chapter4_Supplemental_Table_4.xlsx
- He_chapter4_Supplemental_Table_5.xlsx
- He_chapter4_Supplemental_Table_6.xlsx
- He_chapter4_Supplemental_Table_7.xlsx
- He_chapter4_Supplemental_Table_8.xlsx
- He_chapter4_Supplemental_Table_9.xls

ACKNOWLEDGEMENTS

My voyage to PhD is impossible without the generous supports from many people. Firstly, the help from my committee members (Bing Ren, Wei Wang, Kun Zhang and Vineet Bafna) is tremendous and I am especially grateful to their constructive advice and valuable time. I would also like to thank the members of the Ecker lab, with whom I have endless fun and encouragement. I owe countless beers to Matt Schultz and Bob Schmitz for their guiding me to the right track at the begining of my graduate school, kindly sharing their insights and experiences of science, and being great friends. Finally, I am extremely grateful to my advisor, Joe Ecker, who welcomed me to the lab, taught me how to be a real scientist, provided me the resource and opportunity to work on cutting-edge science, and backed me up when I experienced hard times.

Chapter 2, in full, is a reprint of the material as it appears in Nature 2015. Matthew D. Schultz, Yupeng He, John W.Whitaker, Manoj Hariharan, Eran A. Mukamel, Danny Leung, Nisha Rajagopal, Joseph R. Nery, Mark A. Urich, Huaming Chen, Shin Lin, Yiing Lin, Bing Ren, Terrence J. Sejnowski, Wei Wang, Joseph R. Ecker. Human Body Epigenome Maps Reveal Noncanonical DNA Methylation Variation. Nature. 523(7559):212-216, 2015 Jul. https://www.nature.com/nature/journal/v523/n7559/full/nature14465.html The dissertation author was a co-primary investigator and author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in Proceedings of the National Academy of Sciences 2017. Yupeng He, David U. Gorkin,

Diane E. Dickel, Joseph R. Nery, Rosa G. Castanon, Ah Young Lee, Yin Shen, Axel Visel, Len A. Pennacchio, Bing Ren, and Joseph R. Ecker. Improved regulatory element prediction based on tissue-specific local epigenomic signatures. Proceedings of the National Academy of Sciences. 28;114(9):E1633-1640, 2017 Feb. http://www.pnas.org/content/114/9/E1633 The dissertation author was a primary investigator and author of this paper.

Chapter 4, in full, is a reprint of a manuscript to be submitted to Nature. Yupeng He, Manoj Hariharan, David U. Gorkin, Diane E. Dickel, Chongyuan Luo, Rosa G. Castanon, Joseph R. Nery, Ah Young Lee, Brian A. Williams, Diane Trout, Henry Amrhein, Rongxin Fang, Huaming Chen, Bin Li, Axel Visel, Len A. Pennacchio, Bing Ren and Joseph R. Ecker. Dynamic methylome remodeling throughout mammalian fetal development. In Preparation. The dissertation author was primary investigator and author of this paper.

VITA

2017	Doctor of Philosophy, University of California, San Diego Bioinformatics and Systems Biology
2011	Bachelors of Science, Shanghai Jiaotong University

PUBLICATIONS

He, Yupeng, Manoj Hararan, David Gorkin, Diane E. Dickel, et al. Dynamic methylome remodeling throughout mammalian fetal development. In submission.

He, Yupeng, et al. Improved regulatory element prediction based on tissue-specific local epigenomic signatures. *Proceedings of the National Academy of Sciences* 114, no. 9 (2017): E1633-E1640. http://www.pnas.org/content/114/9/E1633

Ma, Hong, Ryan C. ONeil, Nuria Marti Gutierrez, Manoj Hariharan, Zhuzhu Z. Zhang, **Yupeng He**, Cengiz Cinnioglu et al. Functional Human Oocytes Generated by Transfer of Polar Body Genomes. *Cell Stem Cell* 20, no. 1 (2017): 112-119. http://www.sciencedirect.com/science/article/pii/S1934590916303411

Theunissen, Thorold W.*, Marc Friedli*, **Yupeng He***, Evarist Planet, Ryan C. ONeil, Styliani Markoulaki, Julien Pontis et al. Molecular criteria for defining the naive human pluripotent state. *Cell stem cell* 19, no. 4 (2016): 502-515. http://www.cell.com/cell-stem-cell/abstract/S1934-5909(16)30161-8

Kawakatsu, Taiji, Shao-shan Carol Huang, Florian Jupe, Eriko Sasaki, Robert J. Schmitz, Mark A. Urich, Rosa Castanon, Joseph R. Nery, Cesar Barragan, **Yupeng He** et al. Epigenomic diversity in a global collection of Arabidopsis thaliana accessions. *Cell* 166, no. 2 (2016): 492-505. http://www.sciencedirect.com/science/article/pii/S0092867416308522

Eric M. Scott, Anason Halees, Yuval Itan, Emily G. Spencer, **Yupeng He**, Mostafa Abdellateef, et al. Capture of Greater Middle Eastern Genetic Variation Enhances Disease Gene Discovery. *Nature Genetics* (2016) https://www.ncbi.nlm.nih.gov/pmc/articles/pmid/27428751/

Matthew D. Schultz*, **Yupeng He***, et al. Human Body Epigenome Maps Reveal Noncanonical DNA Methylation Variation. *Nature* (2015), 523, 212216. http://www.nature.com/nature/journal/v523/n7559/full/nature14465.html

He, Yupeng, Joseph R. Ecker. Non-CG Methylation in the Human Genome. *Annual Review of Genomics and Human Genetics* 16.1 (2015). (An invited review that is not peer reviewed). http://www.annualreviews.org/doi/full/10.1146/annurev-genom-090413-025437

Wu, Jun, Daiji Okamura, Mo Li, Keiichiro Suzuki, Chongyuan Luo, Li Ma, **Yupeng He** et al. An alternative pluripotent state confers interspecies chimaeric competency. *Nature* 521, no. 7552 (2015): 316-321. http://www.nature.com/nature/journal/v521/n7552/abs/nature14413.html

Zhang, Yizhe, **Yupeng He**, Guangyong Zheng, and Chaochun Wei. MOST+: A de novo motif finding approach combining genomic sequence and heterogeneous genome-wide signatures. *BMC genomics* 16, no. 7 (2015): S13. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4474412/

Ma, Hong, Robert Morey, Ryan C. O'Neil, **Yupeng He**, Brittany Daughtry, Matthew D. Schultz, Manoj Hariharan et al. Abnormalities in human pluripotent cells due to reprogramming mechanisms. *Nature* 511, no. 7508 (2014): 177-183. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4898064/

Buganim, Yosef, Styliani Markoulaki, Niek van Wietmarschen, Heather Hoke, Tao Wu, Kibibi Ganz, Batool Akhtar-Zaidi, **Yupeng He**, et al. The developmental potential of iPSCs is greatly influenced by reprogramming factor selection. *Cell Stem Cell* 15, no. 3 (2014): 295-309. http://www.cell.com/cell-stem-cell/abstract/S1934-5909(14)00299-9

Castellana, Natalie E., Zhouxin Shen, **Yupeng He**, Justin W. Walley, Steven P. Briggs, and Vineet Bafna. An automated proteogenomic method uses mass spectrometry to reveal novel genes in Zea mays. *Molecular & Cellular Proteomics* 13, no. 1 (2014): 157-167. http://www.mcponline.org/content/13/1/157.long

Schmitz, Robert J*, **Yupeng He***, Oswaldo Valds-Lpez, Saad M. Khan, Trupti Joshi, Mark A. Urich, Joseph R. Nery et al. Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population. *Genome research* 23, no. 10 (2013): 1663-1674. http://genome.cshlp.org/content/23/10/1663

Woo, Sunghee, Seong Won Cha, Gennifer Merrihew, **Yupeng He**, Natalie Castellana, Clark Guest, Michael MacCoss, and Vineet Bafna. Proteogenomic database construction driven from large scale RNA-seq data. *Journal of proteome research* 13, no. 1 (2013): 21-28. https://www.ncbi.nlm.nih.gov/pmc/articles/pmid/23802565/

He, Yupeng, Yizhe Zhang, Guangyong Zheng, and Chaochun Wei. CTF: a CRF-based transcription factor binding sites finding system. *BMC genomics* 13, no. Suppl 8 (2012): S18. https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-13-S8-S18

* indicates co-authorship

ABSTRACT OF THE DISSERTATION

Beyond DNA sequence: Exploring and Exploiting Mammalian DNA Methylation

by

Yupeng He

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California, San Diego, 2017

Professor Joseph R. Ecker, Chair
Professor Wei Wang, Co-Chair

Cytosine DNA methylation (mC) is a chemical modification prevalent in mammalian genome and it plays important roles in transcriptional regulation, development and cell differentiation. Recent studies reveal that mC affects how DNA is interpreted, like an additional information layer on top of the genetic code. Results of both *in vitro* and *in vivo* experiments demonstrate that mC is influential on the binding affinity of a number of transcription factors. Furthermore, targeted addition/removal of mC was shown to modulate gene transcription. In addition, while mC was thought to be stable chemical decoration on DNA, it can be dynamically added or removed during biological processes such as cell differentiation, and its distribution is distinct in different cell types and tissues. Given mC's potential functional impact and cell/tissue specificity, systematically profiling mC across a variety of cell types and tissues is essential for understanding its biological significance. To fill this gap, I first worked with colleagues to dissect the mC landscape of 18 human tissue types from 4 individual. We systematically compared the mC distribution in these human tissues and identified over a million differentially methylated regions, which are strongly overlapped with tissue-specific regulatory DNA elements. The dataset serves as the mC state baseline of normal human tissues. In my second thesis project, I exploited the mC information and developed a computational approach called REPTILE to improve the identification of enhancers, the regulatory DNA elements that promote the transcription of their target genes. Finally, I worked with colleagues to investigate the temporal mC regulation in 12 developing mouse fetal tissues. Our results indicate that mC changes dramatically during development primarily at regulatory DNA elements and it shows a trend of demethylation at fetal stages followed by remethylation after birth. I applied REPTILE on this dataset and delineated hundreds of thousands of enhancers related to tissue development.

Chapter 1

Introduction

DNA stores the code of life. The sequence composed of As, Cs, Gs and Ts instructs how and when a cell functions, replicates, interacts with other cells and differentiates to a different cell type. Thousands of cell types exist in the body of mammals but these very distinct cells carry basically identical DNA information. Epigenetic modifications, including chemical decorations on DNA or proteins that DNA wraps around, likely affect how DNA sequence is interpreted by transcription regulators, shape gene expression landscape and drive the cell type diversity.

1.1 DNA methylation

Cytosine DNA methylation is a chemical modification that methyl group is added to the 5th position of DNA base cytosine. The distribution of DNA methylation can be accurately measure at each single cytosine by whole-genome bisulfite sequencing or MethylC-seq[1]. In mammalian genome, this chemical modification occurs at cytosines followed by guanine (CG methylation) as well as at cytosines followed by non-guanine bases (CH methylation; H = A, C or T)[1]. CG methylation is the most prevalent form and it is present in all cell types and tissues. CG methylation is pervasive in DNA and in somatic cells most of CG sites are methylated, except forCG islands, the genomic regions with high CG density[2]. A traditional view of CG methylation is that it is a stable, repressive epigenetic mark, which is responsible for the repression transposable elements[3]. However, several recent surveys of the CG methylation distribution in tissues revealed that CG methylation is highly variable in regulatory DNA elements such enhancers and CG methylation depletion in these regions is associated with the binding of transcriptions factors and enhancer activities[4, 5, 6, 7]. Though whether CG methylation has a causal effect remain debatable, it is informative about gene regulation and cell disease state[7]. Researchers recently peek into the mechanistic aspect of CG methylation by studying its interaction with transcription factors. They discovered that numerous transcription factors are able to recognize CG methylation and their DNA binding affinity can be reduced or promoted by CG methylation[8, 9, 10, 11]. Non-CG methylation, instead, are understudied and were only found in embryonic stem cells, oocytes, brain, heart, skeletal muscle and several adult tissues, though its functional impact remain obscure[12]. Interestingly, the key protein related to Rett syndrome, methyl CpG binding protein 2 (MeCP2), is able to bind at high affinity to CH methylated DNA[13, 14, 15, 16]. CH methylation has also been linked to type 2 diabetes[12].

In mammals, CG methylation is maintained by the activity of DNA methyltransferase 1 (DNMT1) during cell division: the newly synthesized DNA molecules in semiconservative replication contain no DNA methylation and their pairing with template DNA molecules form hemimethylated CG sites, which are targeted and methylated by DNMT1[17]. However, no mechanism is known to maintain non-CG methylation[12]. The preferable substrate of DNMT1 is hemimethylated DNA whereas it has little activity on unmethylated DNA[17, 18]. Adding methyl groups to cytosines on unmethylated DNA (de novo DNA methylation) relies on the enzymatic activity of DNMT3a and DNMT3b[17, 18]. While both DNMT3a and DNMT3b almost exclusively methylate the cytosines on CG context, previous studies showed that both of them are able to methylated cytosines on non-CG context at much lower level[12].DNA methylation can be both actively and passively removed. Active removal involves the activity of Ten-eleven translocation (TET) enzymes[19]. DNA methylation can also be removed passively in DNA replication during cell division[19].

Although DNA methylation was extensively studied in the past few decades and great progress have been made, we still lack a comprehensive set of DNA methylation maps for various tissues and cell types. Furthermore, it remains difficult to interpret DNA methylation variation and use such information to infer the functional readouts of DNA segments. In this dissertation, I will present the results of three research projects, which specifically address the below questions:

- What is the genomic distribution of DNA methylation (CG and non-CG methylation) in normal human tissues?
- What are the potential functional consequences of tissue-specific DNA methylation variation?
- Can DNA methylation be used to improve the annotation of functional DNA

elements, such as enhancers and super-enhancers[20]?

- How is DNA methylation regulated during development especially within regulatory elements?
- What are the functional implication of the developmental DNA methylation dynamics?

1.2 Outline

Chapter 2 describes the DNA methylation landscape of 18 human tissue types from 4 donors. We found the distribution of DNA methylation is distinct in different tissue types. By systematically identifying differential methylation, over a million genomic regions were pinpointed. These regions are strongly enriched for TF binding motifs and are significantly overlapped with distal transcriptional regulatory elements. In addition, we found the presence of CH methylation at various levels in almost all human tissues. CH methylation is tissue-specifically enriched in gene bodies and is associated with transcription repression. Interestingly, CH methylation is abundant in the bodies of genes that escape X chromosome inactivation. Finally, we found allele-specific DNA methylation, which were linked to allele-specific gene transcription.

In Chapter 3, we present a computational algorithm, REPTILE, which integrates DNA methylation and histone modification data to precisely delineate the location of enhancers. We show that REPTILE outperforms then existing approaches in both accuracy and resolution. REPTILE best predicts the *in vivo* enhancer activity of DNA elements that were experimentally validated. In addition, the location of enhancer predictions from REPTILE is nearer to open chromatin compared to other methods. Thus, we expect REPTILE will be a useful tool for annotating the regulatory landscape of the numerous cell types and tissues.

In Chapter 4, we describe a study about the spatiotemporal distribution of DNA methylation in developing mouse embryo. In this study, we profiled the DNA methylation landscape of 12 mouse tissue types from embryos of 8 fetal developmental stages. Close to 2 million regions were identified as showing differentially methylation. Using REPTILE, we integrated DNA methylation and histone modification data to generate enhancer annotation for each tissue at each fetal stage. Interestingly, these regions predominantly lose CG methylation during fetal development, whereas the trend is reversed after birth. The CG methylation dynamics are closely associated with enhancer activity. In addition to CG methylation, during development, CH methylation is accumulating in almost all tissues in the bodies of many genes that encode TFs related to tissue development, and it is associated with their transcription repression at later fetal development stages. The epigenome maps of developing mouse tissues serve as a valuable resource for studying not only the dynamic transcription regulation in mammalian embryo but also the origin of human birth defects.

1.3 References

[1] R. Lister, M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, J. R. Nery, L. Lee, Z. Ye, Q. M. Ngo, L. Edsall, J. Antosiewicz-Bourget, R. Stewart, V. Ruotti, A. H. Millar, J. A. Thomson, B. Ren, and J. R. Ecker. Human dna methylomes at base resolution show widespread epigenomic differences. *Nature*, 462:315–322, 2009.

[2] Peter a. Jones. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13(7):484–492, 2012. ISSN 1471-0056. doi: 10.1038/nrg3230. URL http://dx.doi.org/10.1038/nrg3230.

[3] A. Bird. Dna methylation patterns and epigenetic memory. *Genes Dev*, 16:6, 2002.

[4] Ryan Lister, Eran A. Mukamel, Joseph R. Nery, Mark Urich, Clare A. Puddifoot, Nicholas D. Johnson, Jacinta Lucero, Yun Huang, Andrew J. Dwork, Matthew D. Schultz, Miao Yu, Julian Tonti-Filippini, Holger Heyn, Shijun Hu, Joseph C. Wu, Anjana Rao, Manel Esteller, Chuan He, Fatemeh G. Haghighi, Terrence J. Sejnowski, M. Margarita Behrens, and Joseph R. Ecker. Global epigenomic reconfiguration during mammalian brain development. *Science*, Jul 2013.

[5] Michael J. Ziller, Hongcang Gu, Fabian Müller, Julie Donaghey, Linus T-Y Tsai, Oliver Kohlbacher, Philip L. De Jager, Evan D. Rosen, David A. Bennett, Bradley E. Bernstein, Andreas Gnirke, and Alexander Meissner. Charting a dynamic dna methylation landscape of the human genome. *Nature*, 500:477–481, Aug 2013.

[6] Gary C. Hon, Nisha Rajagopal, Yin Shen, David F. McCleary, Feng Yue, My D. Dang, and Bing Ren. Epigenetic memory at embryonic enhancers identified in dna methylation maps from adult mouse tissues. *Nat Genet*, 45: 1198–1206, Oct 2013.

[7] Dirk Schübeler. Function and information content of DNA methylation. 2015. doi: 10.1038/nature14192.

[8] S. Domcke, A. F. Bardet, P. Adrian Ginno, D. Hartl, L. Burger, and

D. Schubeler. Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature*, 528(7583):575–579, Dec 2015.

[9] J. Wan, Y. Su, Q. Song, B. Tung, O. Oyinlade, S. Liu, M. Ying, G. L. Ming, H. Song, J. Qian, H. Zhu, and S. Xia. Methylated cis-regulatory elements mediate KLF4-denpendent gene transactivation and cell migration. *Elife*, 6, May 2017.

[10] RonanC. O'Malley, Shao-shanCarol Huang, Liang Song, MathewG. Lewsey, Anna Bartlett, JosephR. Nery, Mary Galli, Andrea Gallavotti, and JosephR. Ecker. Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell*, 165(5):1280–1292, 2016. ISSN 00928674. doi: 10.1016/j.cell.2016.04.038. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867416304810.

[11] H. Zhu, G. Wang, and J. Qian. Transcription factors as readers and effectors of DNA methylation. *Nat. Rev. Genet.*, 17(9):551–565, 08 2016.

[12] Y. He and J. R. Ecker. Non-CG Methylation in the Human Genome. *Annu Rev Genomics Hum Genet*, 16:55–77, 2015.

[13] J. U. Guo, Y. Su, J. H. Shin, J. Shin, H. Li, B. Xie, C. Zhong, S. Hu, T. Le, G. Fan, H. Zhu, Q. Chang, Y. Gao, G. L. Ming, and H. Song. Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. *Nat. Neurosci.*, 17(2):215–222, Feb 2014.

[14] L. Chen, K. Chen, L. A. Lavery, S. A. Baker, C. A. Shaw, W. Li, and H. Y. Zoghbi. MeCP2 binds to non-CG methylated DNA as neurons mature, influencing transcription and the timing of onset for Rett syndrome. *Proc. Natl. Acad. Sci. U.S.A.*, 112(17):5509–5514, Apr 2015.

[15] Harrison W. Gabel, Benyam Kinde, Hume Stroud, Caitlin S. Gilbert, David a. Harmin, Nathaniel R. Kastan, Martin Hemberg, Daniel H. Ebert, and Michael E. Greenberg. Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. *Nature*, 2015. ISSN 0028-0836. doi: 10.1038/nature14319. URL http://www.nature.com/nature/journal/vaop/ncurrent/full/nature14319.html?WT.ec{_}id=NATURE-20150312{#}affil-auth.

[16] S. Lagger, J. C. Connelly, G. Schweikert, S. Webb, J. Selfridge, B. H. Ramsahoye, M. Yu, C. He, G. Sanguinetti, L. C. Sowers, M. D. Walkinshaw, and A. Bird. MeCP2 recognizes cytosine methylated tri-nucleotide and dinucleotide sequences to tune transcription in the mammalian brain. *PLoS Genet.*, 13(5):e1006793, May 2017.

[17] J. A. Law and S. E. Jacobsen. Establishing, maintaining and modifying dna methylation patterns in plants and animals. *Nat Rev Genet*, 11:204–220, Mar 2010.

[18] Bernard H Ramsahoye, Detlev Biniszkiewicz, Frank Lyko, Victoria Clark, Adrian P Bird, and Rudolf Jaenisch. Non-cpg methylation is prevalent in embryonic stem cells and may be mediated by dna methyltransferase 3a. *Proceedings of the National Academy of Sciences*, 97(10):5237–5242, 2000.

[19] Xiaoji Wu and Yi Zhang. Tet-mediated active dna demethylation: mechanism, function and beyond. *Nature Reviews Genetics*, 2017.

[20] Denes Hnisz, Brian J. Abraham, Tong Ihn Lee, Ashley Lau, Violaine Saint-André, Alla A. Sigova, Heather A. Hoke, and Richard A. Young. Super-enhancers in the control of cell identity and disease. *Cell*, 155:934–947, Nov 2013.

Chapter 2

Human Body Epigenome Maps Reveal Noncanonical DNA Methylation Variation

2.1 Summary

Understanding the diversity of human tissues is fundamental to disease and requires linking genetic information, which is identical in most of an individuals cells, with epigenetic mechanisms that could play tissue-specific roles. Surveys of DNA methylation in human tissues have established a complex landscape including both tissue-specific and invariant methylation patterns[1, 2]. Here we report high coverage methylomes that catalogue cytosine methylation in all contexts for the major human organ systems, integrated with matched transcriptomes and genomic sequence. By combining these diverse data types with each individuals

phased genome[3], we identified widespread tissue-specific differential CG methylation (mCG), partially methylated domains, allele-specific methylation and transcription, and the unexpected presence of non-CG methylation (mCH) in almost all human tissues. mCH correlated with tissue-specific functions, and using this mark, we made novel predictions of genes that escape X-chromosome inactivation in specific tissues. Overall, DNA methylation in multiple genomic contexts varies substantially among human tissues.

2.2 Main text

To better understand the variability of DNA methylation across human tissues, we obtained post-mortem samples of 18 tissue types from 4 individuals (5 singletons, 8 duplicates, and 5 triplicates; Figure 2.1a; Methods; Supplementary Table 1) and performed deep transcriptome (36 mRNA-seq samples; 120-475 million reads per sample), base-resolution methylome (36 MethylC-seq[4] samples; 30x-80x genome coverage per sample), and genome sequencing (4 whole genome sequences; 20x-45x genome coverage per sample). We focused our initial analysis on cytosines in the CG context and used a previously published method[1] to identify differential methylation (Methods). We found that 15.4% (4,073,896 of 26,474,560 sites tested) of CG sites in these experiments are strongly differentially methylated (DMS; minimum methylation difference 0.3; Figure 2.5a), which is similar to a previous study[1]. To identify differentially methylated regions (DMRs), we combined sites within 500bp of one another and found 1,198,132 DMRs. Even with these stringent criteria, 719,837 (60.1%) of the DMRs we identified were novel[1, 5].

As expected, hypomethylation at DMRs correlated with tissue-specific functions[1, 6]. For example, strongly hypomethylated DMRs in aorta overlap with aorta-specific super enhancers[7] around *MYH10*, a gene involved in blood vessel function[8] (Figure 2.1b). To further validate our DMRs, we performed hierarchical clustering on their weighted methylation levels[9] (Methods; Figure 2.1c; Figure 2.5b, c). Tissues that were part of the same organ system clustered together (e.g., heart and muscle tissues). We compared these results to a clustering of differentially expressed genes identified in the transcriptomes and found a similar separation of organ systems (Methods; Figure 2.1d; Figure 2.5d). Furthermore, GREAT[10] analysis on the most hypomethylated tissue-specific DMRs revealed many tissue-specific functions (Figure 2.5e, f; Methods; Supplementary Table 2-3).

To examine the relationship between methylation and transcription, we correlated the methylation levels of DMRs and the expression of the closest genes (Figure 2.2a; Figure 2.6a, b; Methods). As expected, methylation in DMRs had a negative correlation with expression, and this correlation grew stronger closer to the transcription start site (TSS). The strongest negative correlation was not in gene promoters but downstream of the promoter up to 8kb away (intragenic vs. promoter median spearman correlation coefficient (SCC) difference -0.12; Mann-Whitney P-value 6.7e-17; Figure 2.2a). This analysis shows that transcription is strongly associated with intragenic DMRs in the tissues we examined, extending similar observations in cancer methylomes[11]

These intragenic methylation differences have previously been hypothesized to mark intragenic CG islands (CGIs) or CGI shores[5, 12, 13, 14]. However, only a small fraction of intragenic DMRs fell in these features (19%; Figure 2.6c).

In addition, predicted enhancers and putative promoters only accounted for 23% and 22% of intragenic DMRs, respectively, suggesting that the remaining DMRs, which we call undefined intragenic DMRs (uiDMRs), represent an unrecognized set of functional elements (35%; Figure 2.6c; Methods). The methylation level of these uiDMRs correlated strongly with the expression of the genes containing them. To examine their regulatory potential, we plotted their histone modification profiles (H3K4me1, H3K4me3, H3K27ac, H3K9me3, H3k27me3 and H3K36me3) derived from the same tissue samples[15] and found five classes: weak enhancer, promoter-proximal, transcribed, poised enhancer and unmarked. (Figure 2.6d-h, Figure 2.7a, b; Methods). Classes with strong, active histone modifications were moderately negatively correlated with expression (weak enhancer and proximal promoter uiDMRs; median SCC -0.31 and -0.16, respectively); whereas, uiDMRs with less active histone modifications exhibited a weak negative correlation (transcribed and poised enhancer uiDMRs). Notably, the correlation between expression and methylation at promoter-proximal uiDMRs was as strong as the correlation with intragenic DMRs that overlapped strong promoters (Figure 2.8; Methods), indicating that intragenic promoter and promoter-proximal sequences are more predictive of changes in methylation than those enriched for enhancer-like chromatin modifications.

In contrast, unmarked uiDMRs showed a weakly positive correlation with expression (Figure 2.8d). Interestingly, we found many of the motifs in tissue-specific uiDMRs were present in tissue-specific enhancers (e.g., *HNF4a*[16] in liver-specific uiDMRs), suggesting that these DMRs are tissue-specific regulatory elements (Methods; Supplementary Table 4-5). Recently, hypomethylated regions that appear inactive in adult tissues but active during fetal development were identified in mice[6]. We examined the DNase I hypersensitivity profiles of unmarked uiDMRs in matched fetal tissues[17] and found an enrichment of hypersensitivity (Figure 2.9; Supplementary Table 6), suggesting that hypomethylation of inactive DMRs can be maintained at regions active earlier in development.

We next examined whether variation in methylation is associated with genetic variation across individuals, which has not been widely characterized in healthy primary tissues or using whole genome bisulfite sequencing[18, 19].To identify individual-specific DMRs, we used a method[20] that is sensitive to these differences unlike the methodology employed above (Methods). We first restricted our analysis to our triplicated samples and ranked DMRs by a tissue-specific methylation outlier score (MOS). We found a 1.6-fold enrichment of SNPs associating with methylation changes in the top 2,500 MOS ranked DMRs in all tissues (Methods). We then used the Epigram pipeline[21] to predict tissue-specific methylation from DNA motifs in these DMRs and found them highly predictive (average area under the curve (AUC) 0.79; Methods). These full models used an average of 156 motifs; however, an average AUC of 0.74 was achieved using only 20 core TF motifs per tissue.

We then identified groups of corresponding motifs by clustering the sets of tissue-specific motifs (Methods). The motif groups were clustered by their tissue hypo- and hypermethylation specificities (Figure 2.2b). 42 of 95 motifs only had hypomethylation specificity; for example, MEIS, which is involved in heart development[22], is hypomethylated in left ventricle, right atrium and right ventricle. We also identified 34 motifs with tissue-dependent methylation specificity.

Three of these motifs match TF families (FOX, HOX and GATA) and are most significantly enriched in hypomethylated regions, suggesting they are primarily involved in regulating hypomethylation.

Mammalian cells have high genome-wide levels of mCG, with the exception of a cultured human fetal fibroblast cell line (IMR90)[4], cancer cells[23, 24] and placenta (PLA)[25]. Surprisingly, large regions of the pancreatic methylomes (PA-2 and PA-3) were significantly hypomethylated (Figure 2.10a). We developed a method to identify PMDs genome-wide (Supplementary Tables 7-8; Methods) and found pancreatic PMDs were smaller than those in IMR90 and PLA (Figure 2.10b) and covered a smaller fraction of the genome (Figure 2.2c). All pairs of PMDs overlapped significantly indicating that these regions are largely shared (¿40% overlap; P-value ¡ 0.001; Figure 2.10c).

Genes in samples with PMDs are transcriptionally repressed[25, 26], but these regions also show reduced expression in all of the tissues we surveyed whether or not a PMD is present (Figure 2.2d). In both IMR90 and PA-2, these regions showed an enrichment in repressive modifications (H3K27me3 and H3K9me3; median difference 0.025 0.168 RPKM (reads per kilobase per million); Mann-Whitney P-value ¡ 2.51e-161) and a depletion in active modifications (H3K4me1, H3K27ac, and H3K36me3; median difference 0.050 0.012 RPKM; Mann-Whitney P-value ¡ 2.03e-53) compared to shuffled regions (Figure 2.2e, f; Figure 2.10d, e; Methods), which provides a potential mechanism for their repression. To try to account for this global hypomethylation, we plotted the expression levels of *DNMT1*, *DNMT3A*, *DNMT3B* and *DNMT3L* but found no systematic expression difference between samples with and without PMDs (Figure 2.11a-d).

Previous studies have highlighted the existence of methylation outside of the CG context (mCH) in human embryonic stem cells[4], brain[2, 20] and at the promoter of the PGC-1 gene in skeletal muscle[27]. We found evidence for appreciable amounts of mCH in many of these tissues (Figure 2.3a; Figure 2.12a). A 5bp motif split the samples into two groups, one with mCH enriched in a TNCAC motif and another with mCH enriched in an NNCAN motif (where N is any base) (Methods). The TNCAC motif is highly similar to the one previously identified in purified glia (GLA) and neurons (NRN) (TACAC). These motifs are significantly different than the motif found in H1 embryonic stem cells (H1) and induced pluripotent stem cells (TACAG)[4, 26] (Figure 2.3b-d). We quantified the extent of mCH across these samples by plotting the distribution of methylation levels at mCH sites in the 25 samples with a TNCAC motif, which revealed a methylation level similar to that of GLA, NRN and H1 (Figure 2.12b)[4, 20]. Most of the tissue types were consistently enriched for the TNCAC or NNCAN motif, but several (esophagus, lung, pancreas and spleen) had replicates which disagreed, suggesting that mCH is not homogenously distributed across these tissues.

To examine the potential functional effect of mCH in adult tissues, we plotted the distribution of expression levels for various quantiles of gene body mCH as it was previously reported to be positively correlated with expression in H1[4] and negatively correlated with expression in neurons[20]. This analysis revealed a negative correlation between expression and mCH (Figure 2.12c; Methods). Next, we combined our replicates and clustered genes by the patterns of CAS methylation (where S is a G or C) in and around their gene body (Figure 2.3e; Methods). To characterize the genes assigned to each cluster, we performed DAVID functional annotation clustering (Supplementary Table 9; Methods), which revealed several different classes. Clusters 1, 2, 11, 16 and 19 contained genes highly enriched for terms involved in basic cellular processes and had an active methylation state (i.e., hypermethylation in embryonic samples and hypomethylation in tissue and brain samples) across all samples. Clusters 5 and 6 were dominated by terms related to neuronal function and genes in this class were differentially methylated between neurons and glia and have inactive methylation states in other samples (i.e., hypomethylation in embryonic samples and hypermethylation in tissue and brain samples). Cluster 12 was enriched for heart and muscle related terms and its genes had an active methylation state in the three heart tissues as well as a weakly active methylation state in psoas but appeared inactive in other samples. Lastly, cluster 14 possessed an active methylation state in brain and tissue samples but were inactive in embryonic samples. Despite being inactive in the H1 samples, this class of genes was highly enriched for terms related to development.

To better define the transition of mCH motifs over development, we examined the ratio of the methylation level of CAC and CAG (mCAC and mCAG) sites in a variety of differentiated (tissues, NRN, and GLA), embryonic (H1), and embryonic derived cells (neural progenitor cells, NPC; mesendoderm MES; trophoblast-like TRO; mesenchymal stem cells, MSC)[28] samples (Figure 2.3f). With the exception of brain cells, mCH levels drop during differentiation, and the mCAC/mCAG ratios revealed a shift in motif usage across developmental time (Figure 2.3f); although, mCAC and mCAG within the same gene remain tightly correlated in both early embryonic and differentiated tissues (Figure 2.12d, e).

Methylation has previously been shown to be predictive of genes escaping

X chromosome inactivation (XI) in neurons[20]. We investigated this phenomenon in these samples by comparing the promoter mCG and gene body mCH of genes that had previously been identified to escape X chromosome inactivation[29] in 11 tissues with mCH (Figure 2.4a). Female-specific promoter mCG hypomethylation and gene body mCH hypermethylation was present at escapee genes at a similar level as in neurons (Figure 2.13a)[20]. Utilizing these tissue methylomes, gene body mCH was appreciably predictive of biallecially expressed genes (AUC 0.89; Figure 2.13b; Methods). To a lesser extent, we observed female-specific promoter mCH and gene body mCG hypermethylation at escapee genes (Figure 2.13a, c, d). Although female-specific promoter mCG hypomethylation, promoter mCH hypermethylation and gene body mCG hypermethylation are somewhat predictive of XI escapees, female-specific gene body mCH hypermethylation is the most predictive feature of XI escapees (Figure 2.13a, b-e). We detected significant female-specific mCH hypermethylation in 109 of 612 X-linked genes, including 9 genes hypermethylated in all 11 tissues and 72 genes that were significantly hypermethylated in only one tissue (Figure 2.4b). Several genes such as FUNDC1 showed female-specific hypermethylation in several tissues but not in neurons, suggesting a tissue-dependent regulation of the escape from X inactivation.

Allele-specific methylation (ASM) and expression (ASE) may also play a role in the regulation of autosomal genes. To examine these phenomena in human tissues, we combined the RNA-seq and MethylC-seq data sets with phased genotypes for each individual in this study[3, 15] (Figure 2.14a; Methods). Using the triplicate tissue samples (FT, GA, PO, SB, and SX), we identified 8,464 - 48,560 ASM events in the CG context and 48 - 403 ASE genes across these tissues

(Supplementary Table 10-11; Methods). We next looked for ASM events that varied across individuals within a tissue-type (tissue variable) and those that varied across a tissue-type within an individual (individual variable). Of the ASM events that varied, 4.1 7.5% and 54.5 70.0% were individual- and tissue-variable, respectively; whereas, of the ASE events that varied, 0.0 20.0% were individual-variable and 13.3 48.8% were tissue-variable (Figure 2.4c; Methods). Of the ASE events, 38.4 87.4% had an ASM event within 100 kilobases, and of these sites, 76% had an ASM and ASE event that was matched (i.e., a DMR was hypomethylated on the same haplotype as the more highly expressed allele). Furthermore, we found that a larger fraction of ASE genes were observed near ASM events whether or not the events matched (Figure 2.14b, c; Methods). These results demonstrate a link between allele specific methylation and expression in human tissues.

Here we have presented the deepest set of base resolution maps of mCG and mCH to date along with chromatin modification states, haplotype-resolved genome sequences and transcriptional profiles for a large set of human tissues. These data sets allowed us to accurately identify cis-regulatory elements. Additionally, they revealed the existence of mCH genome-wide in a subpopulation of cells from differentiated human tissues, which appears to be repressive. Our analysis of genic mCH indicates that these genes are distinct from those that were previously identified in embryonic stem cells and the brain and showed enrichment for a variety of functions, most surprisingly those involved in development. These analyses raise the intriguing possibility that mCH is utilized in adult stem cells[30] and could help to repress these genes as the cells transition into their differentiated role.

2.3 Acknowledgments

We thank R. J. Schmitz for critical reading of the manuscript. This work is supported by the NIH Epigenome Roadmap Project (U01 ES017166). E.A.M. was supported by National Institute of Neurological Diseases and Stroke grant (K99NS080911). J.R.E was supported by the Gordon and Betty Moore Foundation (GMBF3034). T.J.S and J.R.E are investigators of the Howard Hughes Medical Institute. S.L was supported by NIH fellowship grants F32HL110473 and K99HL119617. The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources that have contributed to the research results reported within this paper. URL: http://www.tacc.utexas.edu. The authors would also like to thank Mid-America Transplant Services, St. Louis, MO for their support of this research effort.

Chapter 2, in full, is a reprint of the material as it appears in Nature 2015. Matthew D. Schultz, Yupeng He, John W.Whitaker, Manoj Hariharan, Eran A. Mukamel, Danny Leung, Nisha Rajagopal, Joseph R. Nery, Mark A. Urich, Huaming Chen, Shin Lin, Yiing Lin, Bing Ren, Terrence J. Sejnowski, Wei Wang, Joseph R. Ecker. Human Body Epigenome Maps Reveal Noncanonical DNA Methylation Variation. Nature. 523(7559):212-216, 2015 Jul. The dissertation author was a co-primary investigator and author of this paper.

2.4 Methods

Additional files referred to throughout these methods can be found here: http://neomorph.salk.edu/SDEC_tissue_methylomes/processed_data/code_data.tar.

gz

2.4.1 Tissue Collection

Adrenal, adipose, thymus, esophagus, vascular, bladder, pancreas, liver, stomach, lung, heart, skeletal muscle, ovary, small bowel, colon, and spleen tissues were obtained from deceased donors at the time of organ procurement at Mid-American Transplant Services (St. Louis, USA) after research consent from family was obtained. Samples were flash frozen with liquid nitrogen. From the following tissues, the luminal epithelial lining was dissected free and flash frozen for this study: esophagus, bladder, stomach, small bowel and colon. For tissue from the aorta, the endothelial layer was dissected free and flash frozen. Genomic DNA Sequencing Library Construction Two g of genomic DNA was extracted from ground, frozen tissue using the DNeasy Blood and Tissue kit (Qiagen, Valencia, CA). The DNA was fragmented with a Covaris S2 (Covaris, Woburn, MA) to 300-400 bp, followed by library preparation using the TruSeq DNA Sample Prep kit (Illumina, San Diego, CA) as per manufacturer's instructions. The library was run on a 2% agarose gel and gel size selected to 400-500bp using the MinElute Gel Extraction kit (Qiagen).

2.4.2 RNA-seq Library Construction

Total RNA from tissues and primary cells was extracted using the RNeasy Lipid Tissue Mini Kit according to protocol (QIAGEN). The mRNA libraries were constructed using the TruSeq RNA Sample Prep Kit V2 (Illumina, San Diego, CA) with 4 g total RNA, according to manufacturer's instructions with modifications to confer strand specificity. The RNA was incubated in the Elute, Prime, Fragment Mix at 94C for 4 min. After first strand synthesis, the product was purified using RNAClean XP beads (Beckman, Brea, CA) as per manufacturer's instructions and eluted in 18 L nuclease free water. Second strand synthesis was performed by adding the RNAClean XP purified product to 2.5 L 10x NEB Buffer 2 (New England Biolabs, Ipswich, MA), 2 L dUTP mix (10mM dATPs, 10mM dGTPs, 10mM dCTPs, and 20mM dUTPs), 0.5 L RNAse H (2 U/L), 1 L DNA Polymerase I (E. coli) (New England Biolabs), and 1 L DTT (100 mM). The 25 L mixture was incubated at 16C for 2.5 hours. The purified ligation products were incubated with 2 L Uracil DNA Glycosylase (Fermentas) before PCR amplification. The completed library was then gel size selected to approximately 350-450 bp using the QIAquick Gel Extraction Kit (QIAGEN). RNA-seq libraries were sequenced using the Illumina HiSeq 2000 (Illumina) instrument as per manufacturers instructions. Sequencing of libraries was performed up to 2 101 cycles. Image analysis and base calling were performed with the standard Illumina pipeline version RTA 2.8.0

2.4.3 MethylC-seq Library Construction

Genomic DNA was extracted from ground, frozen tissue using the DNeasy Blood and Tissue Kit (Qiagen, Valencia, CA). Two g of genomic DNA was spiked with 10 ng unmethylated cl857 Sam7 Lambda DNA (Promega, Madison, WI). The DNA was fragmented with a Covaris S2 (Covaris, Woburn, MA) to 150-200 bp, followed by end repair and addition of a 3 A base. Cytosine-methylated adapters provided by Illumina (Illumina, San Diego, CA) were ligated to the sonicated DNA at 16C for 16 hours with T4 DNA ligase (New England Biolabs). Adapter-ligated DNA was isolated by two rounds of purification with AMPure XP beads (Beckman Coulter Genomics, Danvers, MA). Adapter-ligated DNA (450 ng) was subjected to sodium bisulfite conversion using the MethylCode kit (Life Technologies, Carlsbad, CA) as per manufacturers instructions. The bisulfite-converted, adapter-ligated DNA molecules were enriched by 4 cycles of PCR with the following reaction composition: 25 L of Kapa HiFi Hotstart Uracil+ Readymix (Kapa Biosystems, Woburn, MA) and 5 l TruSeq PCR Primer Mix (Illumina) (50 l final). The thermocycling parameters were: 95C 2 min, 98C 30 sec, then 4 cycles of 98C 15 sec, 60C 30 sec and 72C 4 min, ending with one 72C 10 min step. The reaction products were purified using AMPure XP beads. Up to two separate PCR reactions were performed on subsets of the adapter-ligated, bisulfite-converted DNA, yielding up to two independent libraries from the same biological sample.

2.4.4 SNP Calling

SNPs in each of the four donor genome sequences and the H1 genome were detected as follows. Tissue genome sequence fastq files of four donors were mapped using Bowtie2[31] and its default parameters; whereas, the H1 csfasta files were mapped with Bowtie using these parameters: -C -k 1 -m 1 –best –strata -e 80. The UnifiedGenotyper module of GenomeAnalyzerTK[32] (GATK) version 2.4-7 was used to detect SNPs. Default parameters were used, with -dcov 100. The SNPs detected were compared against the dbSNP database (version 137) for classifying known and novel (individual-specific) SNPs. The confidence score threshold for SNP detection was selected as 30. This is the minimum phred-scaled Q-score threshold, provided as a default parameter for high-confidence SNP detection within the GATK package.

2.4.5 SNP-substituted Reference Genomes

We created four modified reference genomes to account for misclassification of CG sites as mCH sites. To that end, we took high-confidence homozygous SNPs and substituted the SNP bases for a particular individual into the hg19 FASTA file.

2.4.6 MethylC-seq Mapping

Sequencing reads were first trimmed for adapter sequence using Cutadapt[33]. All cytosines in the trimmed reads were then computationally converted to thymines and mapped twice, to a converted forward strand reference and to a converted reverse strand reference. A converted reference is created by replacing all cytosines with thymines (forward strand) or all guanines with adenines (reverse strand) in the reference FASTA file. For mapping we used Bowtie[34] with the following options: “-S”,”-k 1”,”-m 1”,”–chunkmbs 3072”,”–best”,”–strata”,”-o 4”,”-e 80”,”-l 20”, and “-n 0”. Reads were mapped to hg19 reference genome. Any read that mapped to multiple locations was removed and one read from each starting location on each strand from each library was kept (i.e., clonal reads were removed). Note that our pipeline (methylpy) does not currently support paired-end reads. Consequently, for MSC, which only had paired-end reads available, we mapped the first read in each pair to avoid problems in processing overlapping reads.

2.4.7 Methylation Calling

To call methylated sites, we summed the number of reads that supported methylation at a site and the number of reads that did not. We used these counts to perform a binomial test with a probability of success equal to the non-conversion rate, which was determined by computing the fraction of methylated reads in the lambda genome (spiked in during library construction). The false discovery rate (FDR) for a given p-value cutoff was computed by calculating the fraction of sites in the lambda genome that had a p-value less than or equal to the cutoff and then dividing that quantity by the fraction of sites that had a p-value less than or equal to the cutoff across all other chromosomes. Because the p-value distributions for each methylation context are different, this procedure was applied to each three nucleotide context independently (e.g., a p-value cutoff was calculated for CAT cytosines). All methylation data was visualized with the AnnoJ browser[35].

2.4.8 DMR Finding

To find tissue-specific differentially methylated regions (DMRs), we used the method described in Ziller et alnding.[1] Briefly, a beta-binomial distribution was used to model the methylation level of each single CG site in each of the tissues. Then, differentially methylated sites (DMS) were identified if the methylation levels of certain site were significantly different between tissues (p-value ¡= 0.01) and the minimum methylation difference was greater than or equal to 0.3. In the next step, DMSs within 500 bp were merged into DMRs. Lastly, for each DMR, the methylation difference between each of tissue pairs (i.e. pairwise comparisons) was computed and only DMRs that have significant methylation difference (p-value ¡= 0.01) and the methylation difference is greater than or equal to 0.3 in at least one of the pairwise comparisons are retained. The scripts for running this pipeline are included as additional files (DDMR_Identification_CpG_mult.r, DDMR_Identification_RegionAnalysis_mult.r, parallel_run_Ziller.py). The results from this script can be found among the additional files (Ziller_et_al_DMR_finding/DMR_final_with_level.tsv)

To statistically infer DMRs that may vary between individuals (i.e., those DMRs used in Genetic Origins of Methylation Variation), which the above methodology from Ziller et al.[1] does not, we defined a stochastic model of our methylation data sets in which the observed number of reads supporting methylated and unmethylated cytosines at each position in each sample is drawn from a binomial distribution. In each sample at each cytosine in the CG context there is a single parameter, x_n^i, corresponding to the true fraction of methylated alleles in the population, or the methylation level, where i denotes the position of cytosine and n denotes the sample. Our null hypothesis is that the methylation level at this position is equal in all samples ($x_n^i = x^i$ for all n). Our procedure is designed to test whether the observed data are consistent with the null hypothesis, or alternatively if there is a significant deviation from equal methylation levels. To do this, first we compute a goodness-of-fit statistic, s, which was introduced and validated by Perkins et al[36]. Specifically, we arrange the observed data in an Nx2 table, with one row for each of N samples and a column for reads supporting methylated and unmethylated cytosines respectively. The number of observed reads in sample n at position i is o_{nj}^i, where $j = 1$ for methylated reads and $j = 2$ for unmethylated reads. The expected number of reads in sample n with methylation state j under the null hypothesis is e^i_{nj}:

$$e^i_{nj} = (\sum_{m=1}^{N} o^i_{mj})(\sum_{k=1}^{2} o^i_{nk})/M^i$$

where $M^i = \sum_{n=1}^{N} \sum_{k=1}^{2} o^i_{nk}$ is the total number of reads in all samples. The statistic for the goodness of fit is

$$s^i = \sqrt{\frac{1}{2N} \sum_{n=1}^{N} \sum_{j=1}^{2} (o^i_{nj} - e^i_{nj})^2}$$

Next, we simulated read count data under our stochastic model assuming the null hypothesis in the following way: Set all cell counts in the table to zero Randomly select a cell in the table with probability equal to the expected counts divided by the total number of counts in the table $(\frac{e^i_{nj}}{M^i})$. Increment the value in this cell by one. Repeat this procedure M^i times. Finally, calculate the value of the statistic, s^i_{shuff}, for the randomly generated table. This randomization procedure was repeated until we observed 100 iterations with a value of s^i_{shuff} that was at least as extreme as that of the observed data, s, up to a maximum of 3,000 iterations. The p-value at position i was then computed as:

$$p^i = \frac{R^i + 1}{T^i}$$

Where R^i is the number of randomized tables with a statistic greater than or equal to the original tables statistic and T^i is the total number of randomized tables that were computed. Our adaptive permutation procedure ensures that any sites which we may potentially identify as significantly differentially methylated with $p^i < 0.01$ will be sampled 3,000 times. At other sites, we have observed an appreciable number (100) of permutations more extreme than our original test statistic (ss_shuff) and the p-value for these sites will be $p(100+1)/3000 = 0.034$; these sites will therefore not be called as differentially methylated.

To control the false discovery rate (FDR) at our desired rate of 1%, we used a computationally efficient procedure designed for comparing multiple sequential permutation-derived p-values[37]. This procedure is designed to account for the effect of our adaptive permutation procedure on the form of the distribution of p-values. First we generated a histogram of the p-values across all cytosines in CG context. We also calculated the expected number of p-values to fall in a particular bin under the null hypothesis. This expected count is computed by multiplying the width of the bin by the current estimate for the number of true null hypotheses (m_0), which is initialized to the number of tests performed. We then identified the first bin (starting from the most significant bin) where the expected number of p-values is greater than or equal to the observed value. The differences between the expected and observed counts in all the bins up to this point are summed, and a new estimate of m_0 is generated by subtracting this sum from the current total number of tests. This procedure was iterated until convergence, which we defined as a change in the m_0 estimate less than or equal to 0.01. With this m_0 estimate, we were able to estimate the FDR corresponding to a given p-value cutoff by multiplying the p-value by the m_0 estimate (the expected number of positives at that cutoff under the null hypothesis) and dividing that product by the total number of significant tests we detected at that p-value cutoff. We chose the largest p-value cutoff that still satisfied our FDR requirement.

In the next stage of analysis, we combined significant sites (DMSs) into blocks if they were within 250 bases of one another and had methylation changes in the same direction (e.g., sample A was hypermethylated and sample B was hypomethylated at both sites). A sample was considered hypo or hyper methylated if the deviation of observed counts from the expected counts was in the top or bottom 1% of deviations. These residuals were calculated for a position i using the following formula for a given cell in row n and column j of the table:

$$\frac{o^i_{nj} - e^i_{nj}}{\sqrt{e^i_{nj} * (1 - \sum_{m=1}^{N} \frac{e^i_{mj}}{M^i}) * (1 - \sum_{k=1}^{2} \frac{e^i_{nk}}{M^i})}} \tag{2.1}$$

The distinction between hypermethylation and hypomethylation was made based on the sign of the residuals. For example, if the residual for the methylated read count of sample A was positive, it was counted as hypermethylation. Furthermore, blocks that contained fewer than 10 differentially methylated sites were discarded. The DMRs called with this methodology, along with their methylation levels, are in the additional files (https://bitbucket.org/schultzmattd/methylpyandDMR_by_methylpy/DMR_methylpy_matrix).

2.4.9 Benchmark methylpy and other DMR identification methods

To further evaluate the performance of the DMR finder (methylpy) used to find inter-individual DMRs in the section Genetic Origins of Methylation Variation, methylpy was compared with three published DMR finding methods: BSmooth[38], DSS[39] and MOABS[40]. The test was done on methylome data of adrenal gland samples from individual 2 and individual 3 (AD-2 and AD-3) and two aorta samples from the same individuals (AO-2 and AO-3). Data and code for this benchmark can be download from this link (https://drive.google.com/folderview?id=0B1BhFMhr3HTATjdWLUx3d1ZtZHM&usp=sharing). For BSmooth and MOABS, the default settings were used. For DSS, we used 1% FDR cutoff for calling differentially methlyated locus (DMLs). Then DMLs within 300bp were merged and regions containing at least 3 DMLs were called as DMRs. Note that these two parameters are the same as the default settings in MOABS. Only data of chromosome 1 was used in this analysis.

2.4.10 Methylation Levels

Throughout the paper we refer to the methylation levels of regions in various contexts. Unless otherwise noted, these methylation levels are more specifically weighted methylation levels as defined here[9]. Sites predicted to be unmethylated (based on the binomial test) had their methylation level set to zero.

2.4.11 RNA-seq Analysis

RNA-seq mapping was done using Tophat2[41] with default parameters (-r 200, –library-type fr-firststrand) against the human reference genome version hg19. The genomic features were obtained from GENCODE version 14[42]. We used htseq-count to map reads to GENCODE features and generate read counts using (http://www-huber.embl.de/users/anders/HTSeq) using default parameters except -s reverse.

2.4.12 RNA-seq Expression Quantification

In order to quantify expression levels of each of the annotated genomic feature, we implemented the cufflinks module of the Cufflinks suite version 2.1.1[43]. Cufflinks produces FPKM (Fragments per kilobase of feature per million) for each of the annotated features. We used default parameters, except for the use of -upper-quartile-norm option and –max-bundle-frags as 50,000,000. This extreme limit was set to avoid skipping of regions with several fragments. The default value of 1,000,000 would result in several tissue-specific or highly expressed genes to be labeled as HIDATA without an actual FPKM value being reported. Then, we applied quartile normalization to FPKMs, which is described in http://cufflinks.cbcb.umd.edu/manual.html#library_norm_meth. Specifically, we scaled the 75% quartile FPKM of every sample to be the mean 75% quartile FPKM of all samples (i.e., all 36 tissue samples from this study, IMR90, H1, and placenta samples).

2.4.13 RNA-seq Differential Expression Analysis

In order to obtain genes that are differentially expressed across any of the samples in this study, we used htseq-count to map reads to GENCODE features and generate read counts (http://www-huber.embl.de/users/anders/HTSeq) using default parameters except -s reverse. These read counts were tested for differential expression using the quasi-likelihood F-test (glmQLFTest)[44] implemented in edgeR[45]. In contrast to pairwise comparisons (like case vs control or wild-type vs treatment) this test does not require specifying which groups would be different. The set of genes enriched or depleted in one group compared to an average of all other tissues was obtained. An FDR cut-off of 0.05 was used to identify differentially expressed genes.

2.4.14 CG DMR Dendrogram

To create the dendrogram shown in Figure 2.1c, we first used the cmdscale command from R to perform multidimensional scaling and compute the first 15 principal components of the CG DMR methylation level matrix. The percent variance explained from this multidimensional scaling is presented in Figure 2.5c. Next, we used the heatmap.2 function in the R package gplots[46] with the default distance metric, and the Ward hierarchical clustering method on these principal components to generate the dendrogram.

2.4.15 Differentially Expressed Genes Dendrogram

To create the dendrogram shown in Figure 2.1d, we first used the cmdscale command from R to perform multidimensional scaling and compute the first 15 principal components of the RPKM values, which were first normalized by the maximum expression value observed at each locus, from all differentially expressed genes. The percent variance explained from this multidimensional scaling is presented in Figure 2.5d. Next, we used the heatmap.2 function in the R package gplots18 with the default distance metric, and the Ward hierarchical clustering method on these principal components to generate the dendrogram.

2.4.16 Genomic Feature Definitions

Promoters were defined as -1000bp to +300bp region of the transcription start sites of transcripts defined in GENCODE version 1414. Exons and introns were also defined using the GENCODE reference. Putative enhancers were obtained from Leung, Rajagopal, and Jung et al.[15] which were predicted using histone mark profiles. CG islands (CGIs) were downloaded from UCSC genome browser[47]. CGI shores were defined as the 2kb regions extending in both directions from CGIs[13, 5].

2.4.17 DMR Tissue Specificity Determination

To find CG DMRs that are strongly and specifically hypomethylated or hypermethylated in a particular tissue, we ranked tissues by the methylation level of a CG DMR (from lowest to highest). Then, starting from the tissue with the lowest methylation level, we computed the difference in methylation level between adjacent tissues. Next, we identified the largest difference, and if it was greater than or equal to 0.1, we divided the tissues into two groups (i.e., hypomethylated tissues and hypermethylated tissues). If the hypomethylated group contained ten or fewer tissues, the DMR was classified as a tissue-specific, hypomethylated CG DMR in those tissues. If the hypermethylated group had ten or fewer tissues, the CG DMR was classified as a tissue-specific, hypermethylated CG DMR in those tissues. We ignored other CG DMRs (including CG DMRs with difference less than 0.1 between adjacent ranked tissues) were because their tissue specificity was too obscure.

2.4.18 DMR GO Enrichment

We used GREAT[10] with default parameters to find functional terms of genes near CG DMRs as these terms indicate the potential regulatory functions of these CG DMRs. Since too many DMRs can saturate the Hypergeometric Test it uses, we considered at most the top 5,000 DMRs sample-specific DMRs ranked (largest to smallest) by the difference (which has to be greater or equal to 0.1) in methylation level between the hypermethylated and hypomethylated groups as input. Furthermore, we require each of these DMRs to have at least 4 DMSs. We focused on the GO Biological Process and Mouse Phenotype categories and representative results from this analysis are shown in Figure 2.5e and f. The complete results are in Supplementary Tables 2 and 3.

2.4.19 Correlating Methylation States of DMRs with Gene Expression

To compute the correlations shown in Figure 2.2a, we used the nearest gene model to predict the target gene of every DMR (i.e., the gene with the closest transcription start site was predicted as the target gene of a DMR). Then, we computed the Spearman correlation coefficient between the methylation level of that DMR and the expression level of its target gene. Only intergenic hypomethylated DMRs with differentially expressed protein-coding genes as a target were included in this analysis. To understand the role of these DMRs and their association with expression, we grouped them into different categories according the genomic elements they did or did not overlap. Genebody DMRs were defined as those that overlapped gene bodies. Enhancer DMRs were defined as those that overlapped enhancers. Promoter, CGI and CGI shore DMRs were defined as those that overlapped promoters, CGIs, or CGI shores. DMRs not in these categories and lying outside any gene body labeled as intergenic. Finally, undefined intragenic DMRs were those that didnt overlap any of these categories. As a control we shuffled the sample labels of the methylation levels and computed the Spearman correlation coefficients as above, which labeled as shuffled.

2.4.20 Annotating undefined intragenic DMRs and promoter DMRs

We used K-means clustering to cluster the histone modification profiles of undefined intragenic DMRs (uiDMRs). We assigned the strand of target gene to each uiDMR to ensure that the TSSs of target genes were always upstream of uiDMRs and eliminate the possibility that strandedness would affect the clustering. Next, we divided each uiDMR into 10 equally sized bins and we divided the 5kb region on either side of each uiDMR into 100bp bins. We split DMRs into equally sized bins for several reasons. Firstly, DMRs varied in length, and we needed a way of comparing the locations of motifs in different DMRs. Secondly, to estimate and show the location preference of motifs, DMRs needed to be binned in order to get an appreciable number of motif instances falling to fall in each position across a DMR. Finally, we wanted to avoid splitting DMRs into bins with different sizes to keep the analysis unbiased as we did not want to introduce confounding factors like differing bin sizes in a single DMR.

We then created a vector of input-normalized ChIP-seq RPKMs of the six histone marks for each bin. The uiDMRs were then clustered into five groups using these vectors. We labeled these groups as weak enhancer (strong H3K4me1, depleted H3K4me3 and strong H3K27ac), promoter-proximal (near region with strong H3K4me3 and strong H3K27ac and depleted in H3K4me1), transcribed (strong H3K36me3), poised enhancer (strong H3K4me1 and weak H3K27ac) and unmarked (no noticeable active histone marks).

We performed a similar analysis for DMRs that overlapped promoters (i.e., the same fixed window definition previously mentioned). Not all of these regions were active (i.e., marked by H3K4me3 and H3K27ac), so to identify active and inactive promoters we applied K-means clustering to the histone modification profile of promoter DMRs into two categories: strong promoters and unmarked promoters. DMRs in strong promoters showed an H3K4me3 and H3K27ac signal;

whereas, DMRs in unmarked promoters displayed at most a very weak H3K27ac and H3K4me3 signal. Sequence motifs enriched in tissue-specific uiDMRs and tissue-specific enhancers were identified using Homer[48].

2.4.21 DNase I sensitivity analysis

To plot DNase I sensitivity data of fetal tissues in Figure 2.9, we downloaded DNase I data from GEO (GSE18927 and Supplementary Table 6). To profile the DNase I sensitivity of unmarked uiDMRs, we divided each unmarked uiDMR into 10 equally sized bins and the 2.5kb region on either side of each uiDMR into 50bp bins. The DNase I sensitivity RPKMs were calculated for each bin for each unmarked uiDMR, and the values were aggregated to generate the average profile. The same approach was applied to generate the average profiles of DMRs overlapping intragenic enhancers and unmarked uiDMRs with shuffled locations. Only DMRs greater than 200bp in length (i.e., each bin is greater than 50bp) are included in this analysis.

2.4.22 Measuring the genetic origins of DNA methylation

If DNA sequence is involved in regulating DNA methylation we should observed an enrichment of sequence variants where there is epigenomic variation. To rank DMRs by epigenomic variation, we created a tissue-specific methylation outlier score (MOS). The MOS takes advantage of some tissues methylomes being sequences in triplicate and identifies DMRs where one individuals methylation state is divergent from the other two. MOS is calculated as,

$$MOS_i = |\frac{\Delta_{ij} + \Delta_{ik}}{2}| - |\Delta_{jk}| \tag{2.2}$$

, where i, j and k represent the three individuals and Δ_{ij} represents the difference in methylation state scores between individuals j and k. MOS_i represents the degree to which individual i is an outlier at a particular DMR. We subtract $|\Delta_{jk}|$ to account for background level of DNA methylation variability at the DMR. A separate MOS is calculated for each individual at each DMR. Each DMR is assigned its single greatest MOS score and the corresponding individual is considered the outlier. We hypothesized that MOS performs better than standard deviation as it considers the level of similarity between the two concordant replicates. Thus, DMRs where variation might be increased by measurement error are less highly ranked as some measurement errors may be consistent across the samples, and therefore, would increase the variation between the concordant replicates. The motif associated SNPs (maSNP) occurrence in the top 2,500 DMRs ranked by standard deviation was: FT = 1.51; GA = 1.45; PO = 1.61; SB = 1.61; SX = 1.46. These numbers result in an average maSNP occurrence of 1.528. When MOS is used to rank the DMRS the enrichment scores are: FT = 1.58; GA = 1.65; PO = 1.65; SB = 1.60; SX = 1.63. The MOS ranked DMRs result in an average maSNP occurrence of 1.622. Thus, MOS does a better job or ranking DMRs by their enrichment with maSNPs. Further, to determine that maSNP enrichment of DMRs when ranked by MOS was statistically significant we used a Chi-squared test to compare the association between the number of maSNPs and non-maSNPs in a DMR and its SD or MOS rank. To do this, the maSNP and non-maSNP counts were compared between the top MOS ranked 2500 DMRs and the DMRs ranked between 497500 and 500000 (i.e., we constructed a 2x2 table where rows indicated whether or not the DMR was in the top 2500 DMRs and columns indicated whether or not that DMR contained an maSNP). The P-values for maSNP enrichment in the top 2,500 MOS ranked DMRs were: FT = 0.0006811861; GA = 2.443996e-16; PO = 4.2191e-16; SB = 0.00202069 and SX = 6.313224e-08. Thus, demonstrating the significance of the maSNP enrichment in the MOS ranked DMRs. The Chi-squared test of significance was repeated using DMRs ranked by strand deviation: FT = 0.01908347; GA = 0.09873; PO = 6.997994e-07; SB = 0.0003348352 and SX = 0.002674707. In all cases, the P-value was more significant for the MOS ranked DMRs. To evaluate the level of sequence variation at cis-regulatory elements we created sets of DNA motifs that are putatively involved in the tissue-specific regulation of DNA methylation levels at the DMRs. For each tissue we created two de novo motif sets: (i) hypo and (ii) hyper. The tissue-specific de novo motif sets were created using the Epigram pipeline[21] to identify a set of motifs that are discriminative of tissue-specific hypo and hypermethylated regions. Briefly, the Epigram pipeline works as the following: (i) the two sets of sequences (tissue-specific hypo and hypermethylated regions) are balanced so that they have the same distribution of lengths and GC-content; (ii) two de novo motif finding methods, HOMER[48] and its own, are used to identify motifs that are enriched in either set; (iii) a LASSO logistic regression[49] is used to select the motifs that are most discriminative of the two regions; (iv) a Random Forest classifier and 5-fold cross-validation are used to assess the collective ability of the motifs to classify the sequences into hypo or hypermethylated; (v) a second round of feature selection is performed to heuristically select a subset of 20 motifs that has the greatest discrimination power. Thus, the Epigram pipeline identities motifs that are predictive of tissue-specific hypo- and hypermethylation and measures their ability to distinguish the two sets. During the creation of both the de novo and known motif sets it is necessary to have sets of tissue-specific hypo and hypermethylated regions. The tissue-specific hypomethylated regions were taken from the DMR GREAT analysis as previously defined. The set of hypo- and hyper-methylated sequence sets were then balanced so that they were equal in size and had the same distribution of GC-content and region lengths[21]. The number of hypomethylated DMRs for each tissue after sampling ranged from 278 to 15,732 with a mean of 7,307 while the hyper sets ranged from 745 to 12,190 with a mean of 6,028. To create known set of known motifs five motif databases were combined: (i) Transfac[50], (ii) Jaspar[51], (iii) Uniprobe[52], (iv) hPDI[53] and (v) Taipale[54]. We removed known if their name was not listed in GENCODE or they were not annotated with the gene ontology term sequence-specific DNA binding or DNA binding. To make the final set of motifs non-redundant, if there was more than one motif for the same gene, then only the motif with the greatest information content was retained. To calculate motif-breaking cut-offs for the known motifs we created background distribution and took a cut-off that corresponds to a 0.05 P-value. Taking the DMR DNA sequences and shuffling them so that order of nucleotides was randomized created the background distribution sequences. A motif specific background distribution was created by recording the best score of S (see above) in each of the shuffled sequence.

2.4.23 PMD Identification

To identify PMDs, we created a random forest classifier. Random forests are an ensemble machine learning technique (described in detail here (Breiman, L. Random Forests. Machine Learning. 2001)) used for classification. We first visually classified regions on chromosome 22 that we felt were strong candidates as PMDs or non-PMDs (Supplementary Table 7). These regions were then used to train a random forest, which was implemented in the python function RandomForestClassifier from the module sklearn.ensemble[55]. Specifically, we then divided these regions into 10kb nonoverlapping bins and computed the percentiles of the methylation levels at the CG sites within each bin. We divided genome into 10kb non-overlapping bins mainly to reduce the effect of smaller DNA methylation variation. PMDs were first discovered by Lister et al. as large (mean length = 153kb, PMID: 19829295) regions with intermediate methylation level (¡ 70%, PMID: 19829295). Consequently, we chose a large bin size (10 kb) to reduce the effect of methylation variations in smaller scale (such as DMRs). Furthermore, the features (methylation level distribution of CG sites) used in classifier required enough CG sites inside each bin to accurately estimate this distribution, which necessitated a relatively large bin. We excluded 10kb bins with fewer than 10 CG sites because of the same reason mention above: accurately estimating the methylation level distribution of CG sites inside bin required enough number of sites. Therefore, for bins with very few CG sites (¡ 10 here), we were unable to classify them (into PMD or non-PMD). These percentiles were used as features for the random forest. The following arguments were supplied to the Python function: n_estimators = 10000, max_features=None, oob_score=True, compute_importances=True In this procedure, out-of-bag error estimation is used to assess the performance of the classifier. More specifically, when building the classifier, the training data can be bootstrap sampled, which leaves a portion of the data out of the classifiers construction and can later be used to assess the rate at which the classifier is correctly predicting known labels. To assess the performance of our models, we calculated one minus the out-of-bag error rate reported by RandomForestClassifier, which yielded a correct prediction rate of at least 90% (PA-2 - 90.23%, PA-3 - 92.37%, IMR90 - 97.65%, PLA - 92.33%).

2.4.24 Comparing PMDs Called in IMR90, PA-2, PA-3 and Placenta

We used GAT[56] to estimate the significance of the overlap between PMDs in different samples shown in Figure 2.10c. The workspace we used was the human reference genome (hg19) excluding ENCODE blacklisted regions. The options provided to GAT were: –ignore-segment-tracks –num-samples=1000 –bucket-size=10000.

2.4.25 Histone Modification Profiles Across PMDs

To profile the histone marks in PMDs and the surrounding regions shown in Figure 2.2e, f, we divided the 300kb upstream and downstream of each PMD into 10kb bins. The body of PMD was divided evenly into 10 bins. Next, we averaged the input normalized ChIP-seq RPKM for each bin. As a control we shuffled the PMDs and performed the same computation.

2.4.26 Testing Histone Modification Enrichment and Depletion Inside and Outside of PMDs

For each histone mark and separately for each sample (PA-2 and IMR90), we grouped the signal medians displayed in Figure 2.2e, f by whether they were inside or outside of the PMD. Next, we performed a Mann-Whitney test on these groups to estimate the significance of the difference in signal medians inside and outside of PMDs.

2.4.27 mCH Motif Calling

To find the predominant nucleotide context of mCH in each sample, we took the top 800,000 methylated, mCH sites (the least number of sites in the three samples displayed in Figure 2.3b-d) that did not overlap with a heterozygous SNP and input the surrounding (+/- 5bp) nucleotides from the SNP-corrected reference genomes to the seqLogo package[57] in Bioconductor. Distribution of Expression Across mCH Quantiles To examine the correlation between expression and mCH, we binned the expression levels genes into quantiles based on the mCH levels in the tissue where expression was measured. For example, the boxplot in Figure 2.12b labeled 85 contains expression levels from all the genes that were between the 85th and 90th quantile of mCH level. It is important to note that the absolute methylation level for the 85th and 90th quantile will vary from tissue to tissue. We took this approach to account for the differences in cellular heterogeneity between these tissues.

2.4.28 mCH Pattern Clustering

To identify sets of genes that share similar DNA methylation patterns in an unbiased fashion, we applied a procedure that combines dimensional reduction using principal component analysis, followed by clustering[20]. We profiled the methylation level (mCAC/CAC and mCAG/CAG) in gene bodies (TSS-TES) and 5 promoter regions (1 kb upstream of the TSS) within each of 25 samples included in this analysis (collapsed tissue replicates, NRN, GLA, H1 and its derivatives). The methylation level in each sample for each gene was normalized by the average over the genes distal flanking region (50-100 kb upstream of TSS or downstream of TES). Normalized mC/C values were then log-transformed. These data were combined into a matrix of 104 features for each of 17,138 autosomal genes. Any bins with missing data due to insufficient coverage in one of the samples (0.22% of the total) were replaced with the median value of the entire data set. We performed singular value decomposition on this data matrix to identify the linear combinations of methylation features that account for the largest fraction of the total data variance. We retained the top 7 PCs as a low-dimensional representation of robust genomic methylation features, accounting for 70.3% of the total data variance. Next, we used k-means clustering to estimate gene sets with highly similar withinset methylation patterns. We chose to extract k=20 clusters to capture a diverse range of methylation features, while still allowing visualization and statistical enrichment analysis of functional association for each gene set. We repeated the clustering procedure 5 times using random initialization of the cluster centers, choosing as the final estimate the run with the smallest within-cluster sum of distances from each point to the cluster centroid. To display the methylation patterns within these gene clusters in Figure 2.3f, we profiled the methylation level (mCAS/CAS) in bins of size 1 kb starting 100 kb upstream of the TSS and ending 100 kb downstream of the transcription end site (TES). To compare genes with different lengths, we divided each gene body into 10 non-overlapping bins of equal size extending from the TSS to the TES. Methylation levels were normalized by the flanking region as described above. We then linearly interpolated the gene-body mCAS/CAS data at 100 evenly spaced bins within the gene body in order to give roughly equal weight to the gene-body and flanking methylation data. To visualize the heatmaps of mCAS/CAS patterns for each of 17,138 genes, we smoothed and downsampled the genes 40-fold to allow representation of genome-scale features.

2.4.29 CAC and CAG Correlation Analysis

In Figure 2.12c, d we examined the relationship between mCAC and mCAG in the following way. The total methylation level (mCAC/CAC or mCAG/CAG) was calculated within all autosomal gene bodies (from TSS to TES). We excluded genes shorter than 2kb. We computed the Spearman (rank) correlation coefficient between these two methylation levels across all genes. These correlations may be diminished by noise due to sampling a finite set of reads for each gene. To determine the magnitude of this effect, we simulated MethylC-Seq basecalls under the assumption of a perfect rank correlation of the true methylation levels. The rank correlation of the simulated reads provides an upper bound on the level of correlation that could have been observed.

2.4.30 Read Position Methylation Level Biases

It has previously been noted that sequencing biases may erroneously be interpreted as mCH[38, 58]. To test for this possibility, we constructed m-bias plots as described here[38] and found a very slight bias in the methylation level at the beginning of our reads (Figure 2.12f-h). Consequently, we trimmed the first 10 bases of reads in a sample with (PO-2) and without (EG-2) mCH to see if this bias affected our identification of the CAC mCH motif. This analysis revealed that the original and bias-free motifs are highly concordant with the mCH motif becoming slightly stronger in the bias-free sample (Figure 2.12i-l). Given that this gain was so slight, we did not feel it justified discarding roughly 10% of our data, so we proceeded with the untrimmed results.

2.4.31 X Chromosome Inactivation

Gender-specific methylation patterns were examined in 9 pairs of tissue samples from adult male (STL003) and female (STL002), as well as paired neuronal (NeuN+) and glial (NeuN) samples from adult male (55yo) and female (53yo)[20]. For each of the genes assayed here[29], we examined the total mCG/CG within the promoter region, defined to be a 1 kb region ending at the TSS, and the total mCG/CG or mCH/CH within the gene body (TSS to TES). For this analysis, we included 612 X-linked genes that were ¿1 kb in length and met a coverage criterion (¿4000 basecalls at CG and CH positions within the gene body in all 22 samples examined). The heatmap in Figure 2.4b shows the ratio of gene body mCH/CH in female vs. male, without any correction for the non-conversion rate. The black outline in Figure 2.4b indicates genes that were found to be significantly hyper-mCH in female (likelihood ratio test, Yekutieli-Benjamini FDR 0.15), with at least 1.2-fold greater mCH/CH in female vs. male, and with mCH/CH¿0 in the female sample (Fisher exact test, p¡0.01). The likelihood ratio test takes into account the sample-specific bisulfite non-conversion rate for mCH sites, as calibrated using sequencing of unmethylated lambda phage DNA. To assess the relationship between female-specific mCH/CH and escape from X-chromosome inactivation (XCI), we relied on a published survey of expression on the inactive human X-chromosome[29]. That study used rodent/human somatic cell hybrids to assign a XCI score to each gene; 0 corresponds to inactivated genes, 9 to escapees, and intermediate values show varying levels of expression from the inactivated X-chromosome. We used liftOver to match 405 of the surveyed genes to our pool of 612 X-linked genes; this set included 34 escapee genes (XCI=9). The box plot (Figure 2.4b) shows the difference between female and male methylation level for genes ranked according to the X-inactivation status index[29]. For each box, the central black line is the median and the box edges are the 25th and 75th percentiles. We used receiver operating characteristic (ROC) analysis to assess how well female-specific mCH hypermethylation allows discrimination of X-escapee genes (Figure 2.13b). The area under the ROC curve (AUC) is a statistical measure of discriminability, which ranges from 0.5 when little or no discrimination information is present to 1 for perfect discriminability. A similar analysis was done to assess how informative female-specific promoter CG hypomethylation, female-specific promoter mCH hypermethylation and female-specific gene body mCG, respectively, is for predicting X-escapee genes. Results are shown in Figure 2.13c-e.

2.4.32 Haplotype Reconstruction using HaploSeq

First, genotypes for all donors were obtained as above. Next, Hi-C reads and paired-end genome sequencing reads were mapped independently using Novoalign (http://www.novocraft.com) to the donor variant-masked hg19 genome as described above. We mapped the Hi-C reads as single ends and paired them later using in-house scripts. We then performed GATK walkers such as Indel realignment and base recalibration to obtain high quality mapping. Finally, we combined our high-quality genome sequencing and Hi-C reads and performed HaploSeq[3] to obtain higher resolution haplotypes than using Hi-C data alone. We then improved the resolution of our seed haplotype generated by HapCUT[59] using local conditional phasing. Briefly, local conditional phasing is performed by Beagle (v4.0)[60] using all known variants in the population (1000 Genomes dataset, phase1 v3). Using the seed haplotypes generated by HapCUT, Beagle infers the haplotype of unphased gap variants using a Hidden Markov Model. In order for a variant to be conditionally phased, we required a 100% match between the phase status present in the seed haplotype and the phase status predicted by Beagle.

2.4.33 Allele-specific Mapping of methylome data

We first generated modified references for each sample (STL001, STL002, STL003, and STL011) to avoid biasing mapping towards reads containing the hg19 reference variant. To this end, we used the SNP calls described above and identified high quality SNPs by recalibrating variants using the default parameters of variant recalibration (GATK) (2) and only genotypes of highest quality (100% confidence calls by GATK) were used for downstream analyses. We masked any heterozygous

SNP with a PASS by replacing them with an N and replaced any homozygous SNP with the appropriate variant. Using these references, we remapped our methylome data with Bowtie2[31] as this aligner allows for alignment to sequences containing Ns using the default settings with the following modifications: "-k 2","–np 0".

2.4.34 Assigning methylome reads to alleles

Mapped methylome reads were assigned to alleles based on base calls on reads that overlapped phased heterozygous SNPs. For reads overlapping multiple phased heterozygous SNPs, they were assigned to allele with support from majority of phased heterozygous SNPs and reads were discarded if two alleles were with equal support. To assign reads to a particular allele, we used the scripts assign_read_to_allele_WGBS_se.pl found in the assign_reads folder (additional files).

2.4.35 Allele-specific methylation analysis

Methylome reads assigned to each allele, were then processed in the same way as that we used for whole sample, which is described above. Then, by comparing methylomes of two alleles, DMRs (i.e. allele-specific methylation (ASM) events) were called using the same approach as described above. We also separated ASM events that were caused by changing one of the alleles cytosine context (i.e., it occurred in one of the two bases following the methylated cytosines) and those that did not. Furthermore, we required that each allele was covered by at least 10 reads. The sequence context of ASM may differ in two alleles and only ASM events that contain CG site(s) in at least one allele were included in following analysis.

2.4.36 Aligning RNA-seq reads to alleles

List of genes showing allele-specific expression in each tissue sample was obtained from Leung, Rajagopal, and Jung et al.[15]. Specifically, For RNA-seq data of all tissue samples, the paired-end reads were mapped using Novoalign to a variant masked transcriptome genome, which was constructed using Useq software based on Gencode annotation (hg18). The mapped reads were assigned alleles according to the sequence match in each variant between two alleles. Then, for each allele, duplicate reads were considered as PCR duplicates and removed with Picard. To determine whether removing duplicate reads in RNA-seq datasets is appropriate during downstream analysis, we investigate the distribution of duplicate reads in terms of gene expression levels. If the duplicate reads are biased to the highly expressed genes the duplicate reads reflect gene expression levels. If not, the duplicate reads can be considered as PCR duplicate reads. We observed that the samples containing high duplicate reads showed uniformly distributed duplicate reads regardless of gene expression levels (data not shown), indicating that the duplicate reads contain a lot of PCR duplicate reads. To avoid any statistical bias during downstream analysis we decided to remove duplicate reads across whole samples. Although reads were aligned to variant-masked genome, there are still others biases favoring either of alleles. First, to reduce the effect of the mappability bias, we aligned simulated reads spanning surrounding variants location and then checked if one allele was favored than the other. If more than 5% reads were mapped to one allele than the other, those variant loci were removed as they are likely to subject inherent mapping bias. Second, to reduce the effect of copy number variation and allelically biased copy number variable regions on allelic analysis, we compared the coverage between two alleles based on WGS data. Any variant that had more than three standard deviations above the mean coverage of each haplotype was excluded. Any variant showing biased WGS coverage between two alleles was also excluded (binomial test p-value less than 0.05 after Benjamini correction). Lastly, we remove heterozygous variants that were erroneously called during genotyping. The probability of each called heterozygous variant that was actually homozygous was calculated from the likelihood of observing the coverage on each allele from whole genome sequencing. Only heterozygous SNPs that had a FDR of less than 0.5% were included in downstream analysis. To identify allelically expressed genes, we performed binomial test (with probability 50% as null hypothesis) on the numbers of aligned reads of two alleles. Only reads spanning exonic regions were counted and only genes containing at least 10 aligned reads were tested. Allelically expressed genes were defined based on 5% FDR cutoff.

2.4.37 Tissue and Individual Variability of Allele-specific Methylation and Expression

We defined an ASM (and ASE) event as individual variable if there was any disagreement across the tissues from a single individual (e.g., FT-1 had an ASM event and SX-1 did not). Similarly, we called a site tissue variable if there was any disagreement across a single tissue from the three individuals (e.g., SB-2 had an ASM event and SB-3 did not).

2.4.38 Association between Allele-specific Methylation and Expression

If there is a strong association between allele-specific methylation (ASM) and allele-specific expression (ASE) events, we should expect more allelic expressed genes rather than bi-allelic expressed genes are proximal to ASM events. To test this, we calculated the fraction of ASE genes and bi-allelically expressed genes that have at least one ASM event within a certain distance. Bi-allelically expressed genes were defined as genes that were covered by at least 10 reads and whose p-values given by binomial test for allelic expression were greater than 0.2. Then, since the distance between genuine ASM and ASE events was unknown, we varied the distance cutoff from 10kb to 100kb. The computation was done for all samples from triplicate tissues and the aggregated the results are shown in Figure 2.14b. Similarly, if ASE is associated with ASM, we should expect more allelic expressed genes can be linked to matched ASM event(s) than matched ASM event(s) with their locations shuffled. Therefore, we computed the fraction of ASE genes that were linked to matched ASM event(s) and matched ASM events but with their locations shuffled. Similar to analysis above, distance cutoff was varied from 10kb to 100kb. The aggregated the results of samples from triplicate tissues are shown in Figure 2.14c.

2.5 Figures

Figure 2.1: **The methylomes and transcriptomes of human tissues. a,** The tissues analyzed in this study. Samples are denoted by the two letter code in parentheses followed by an individual ID. **b,** Browser screenshot of an example DMR. The top track contains gene models. The following four tracks contain green blocks indicating the location of super enhancers, enhancers, and hypomethylated DMRs in aorta, respectively. The remaining tracks display methylation data from each sample. Gold ticks are CG sites with heights proportional to their methylation level. Ticks on the forward and reverse strand are projected upward and downward from the dotted line, respectively.**c-d,** Hierarchical clustering of DMR methylation levels (c) and expression levels of differentially expressed genes (d). Colors indicate organ systems each sample belongs to.

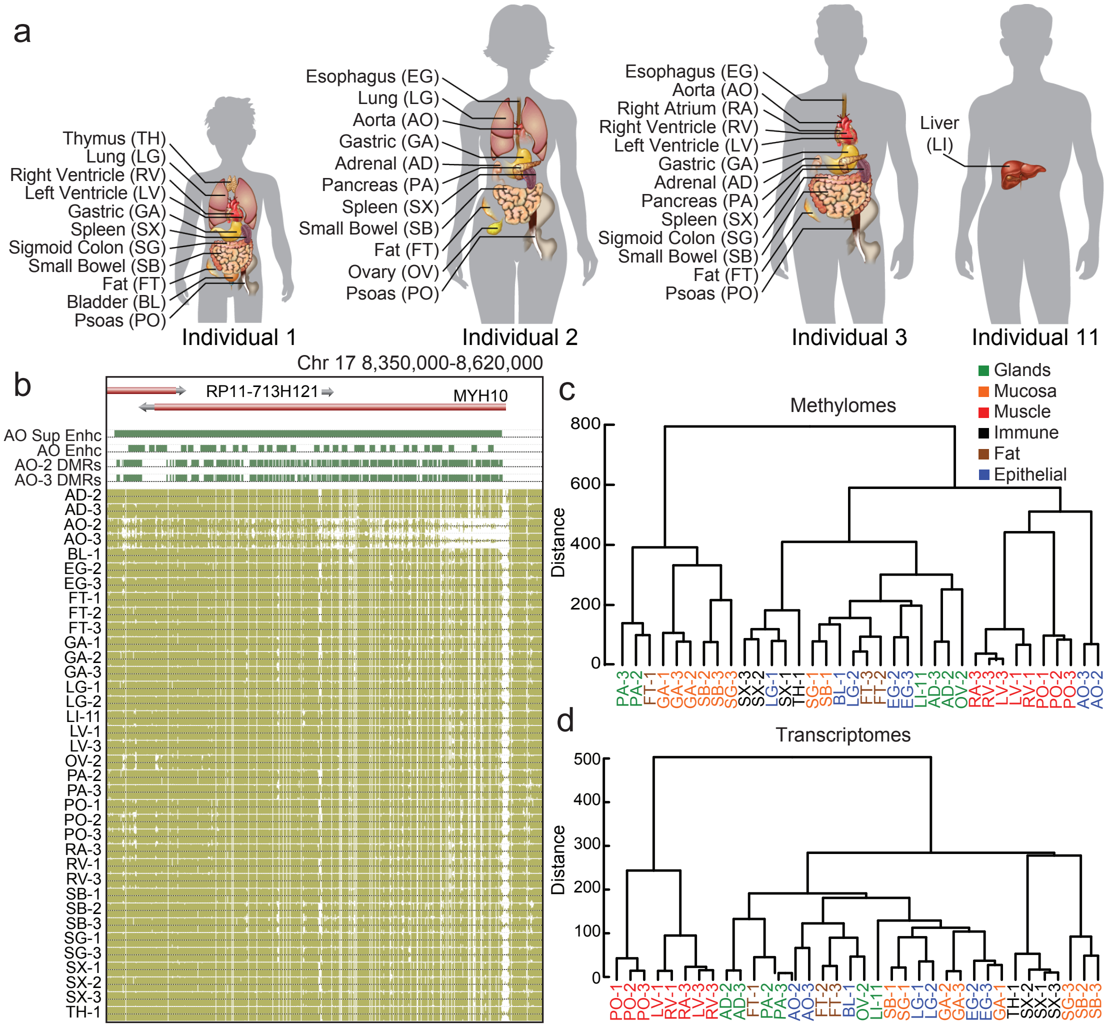

a
Thymus (TH)
Lung (LG)
Right Ventricle (RV)
Left Ventricle (LV)
Gastric (GA)
Spleen (SX)
Sigmoid Colon (SG)
Small Bowel (SB)
Fat (FT)
Bladder (BL)
Psoas (PO)
Individual 1
Esophagus (EG)
Lung (LG)
Aorta (AO)
Gastric (GA)
Adrenal (AD)
Pancreas (PA)
Spleen (SX)
Small Bowel (SB)
Fat (FT)
Ovary (OV)
Psoas (PO)
Individual 2
Esophagus (EG)
Aorta (AO)
Right Atrium (RA)
Right Ventricle (RV)
Left Ventricle (LV)
Gastric (GA)
Adrenal (AD)
Pancreas (PA)
Spleen (SX)
Sigmoid Colon (SG)
Small Bowel (SB)
Fat (FT)
Psoas (PO)
Individual 3
Liver (LI)
Individual 11
b
Chr 17 8,350,000-8,620,000
RP11-713H121
MYH10
AO Sup Enhc
AO Enhc
AO-2 DMRs
AO-3 DMRs
AD-2
AD-3
AO-2
AO-3
BL-1
EG-2
EG-3
FT-1
FT-2
FT-3
GA-1
GA-2
GA-3
LG-1
LG-2
LI-11
LV-1
LV-3
OV-2
PA-2
PA-3
PO-1
PO-2
PO-3
RA-3
RV-1
RV-3
SB-1
SB-2
SB-3
SG-1
SG-3
SX-1
SX-2
SX-3
TH-1
c
Methylomes
🟩 Glands
🟧 Mucosa
🟥 Muscle
⬛ Immune
🟫 Fat
🟦 Epithelial
Distance
800
600
400
200
0
PA-3 PA-2 FT-1 GA-1 GA-3 GA-2 SB-2 SB-3 SG-3 SX-3 SX-2 LG-1 SX-1 TH-1 SG-1 SB-1 BL-1 LG-2 FT-3 FT-2 EG-2 EG-3 LI-11 AD-3 AD-2 OV-2 RA-3 RV-3 LV-3 LV-1 RV-1 PO-1 PO-2 PO-3 AO-3 AO-2
d
Transcriptomes
Distance
500
400
300
200
100
0
PO-1 PO-2 PO-3 LV-1 RV-1 RA-3 LV-3 RV-3 AD-2 AD-3 FT-1 PA-2 PA-3 AO-2 AO-3 FT-2 FT-3 BL-1 OV-2 LI-11 SB-1 SG-1 LG-1 LG-2 GA-2 GA-3 EG-2 EG-3 GA-1 TH-1 SX-2 SX-1 SX-3 SG-3 SB-2 SB-3

Figure 2.2: **DNA methylation and its relationship with gene expression. a,** The mean Spearman correlation coefficient at various distances between the methylation level of autosomal DMRs and the expression of the nearest gene. These correlations are shown for DMRs: overlapping genes (Genebody), overlapping enhancers (Enhancer), overlapping promoters or CpG islands (CGIs) or CGI shores (Promoter, CGI, CGI shore), not overlapping genes (Intergenic) and all remaining DMRs (Undefined). **b,** Heatmap showing each motifs tissue-specific methylation preference. The tissues are colored according to Figure 2.1c, and the ordering is listed at the bottom of the figure. The bar plot at the end of the panel shows the number of times the motif was present in the 20 motif models. **c,** The number of base pairs covered by PMDs in all samples. **d,** The distribution of expression inside and outside of PA-2 PMDs across various samples. Notches indicate a confidence interval estimated from 1,000 bootstrap samples. Each PMD boxplot consists of 3,627 genes and each non-PMD boxplot consists of 22,907 genes. **e-f,** Histone modification profiles in and around PMDs in PA-2 (e) and IMR90 (f).

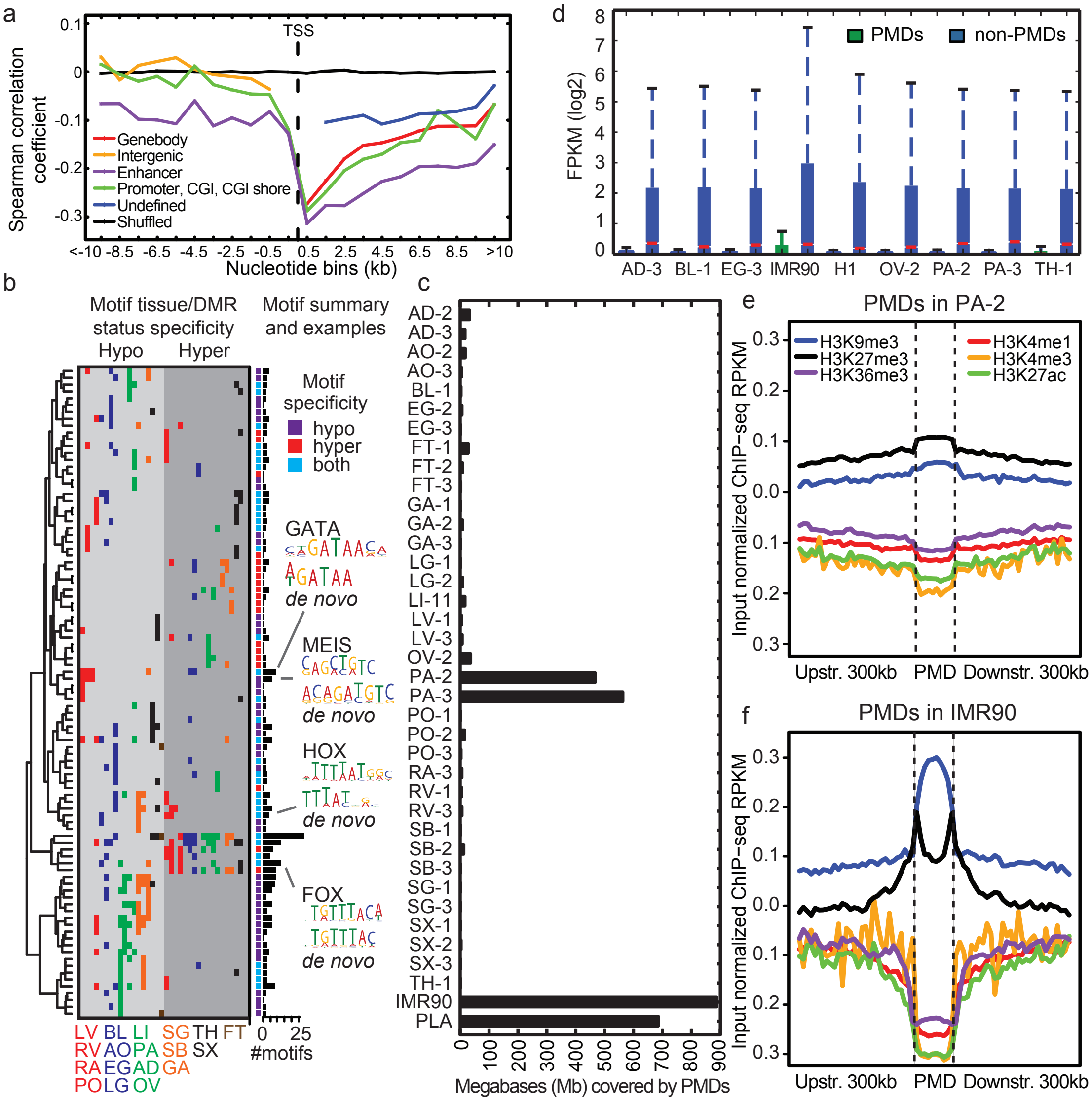

a
Spearman correlation coefficient
0.1
0
-0.1
-0.2
-0.3
TSS
Genebody
Intergenic
Enhancer
Promoter, CGI, CGI shore
Undefined
Shuffled
<-10 -8.5 -6.5 -4.5 -2.5 -0.5 0.5 2.5 4.5 6.5 8.5 >10
Nucleotide bins (kb)
d
FPKM (log2)
PMDs
non-PMDs
AD-3 BL-1 EG-3 IMR90 H1 OV-2 PA-2 PA-3 TH-1
b
Motif tissue/DMR status specificity
Hypo
Hyper
Motif summary and examples
Motif specificity
hypo
hyper
both
GATA
de novo
MEIS
de novo
HOX
de novo
FOX
de novo
LV BL LI SG TH FT
RV AO PA SB SX
RA EG AD GA
PO LG OV
0 25
#motifs
c
AD-2 AD-3 AO-2 AO-3 BL-1 EG-2 EG-3 FT-1 FT-2 FT-3 GA-1 GA-2 GA-3 LG-1 LG-2 LI-11 LV-1 LV-3 OV-2 PA-2 PA-3 PO-1 PO-2 PO-3 RA-3 RV-1 RV-3 SB-1 SB-2 SB-3 SG-1 SG-3 SX-1 SX-2 SX-3 TH-1 IMR90 PLA
0 100 200 300 400 500 600 700 800 900
Megabases (Mb) covered by PMDs
e
PMDs in PA-2
Input normalized ChIP-seq RPKM
H3K9me3
H3K27me3
H3K36me3
H3K4me1
H3K4me3
H3K27ac
Upstr. 300kb PMD Downstr. 300kb
f
PMDs in IMR90
Input normalized ChIP-seq RPKM
Upstr. 300kb PMD Downstr. 300kb

Figure 2.3: **mCH is prevalent in human tissues. a,** The fraction of methylated cytosines in the CH context by sample. **b-d,** Representative mCH motifs from embryonic, (H1; b), tissue (LI-11; c), and brain (NRN; d) samples. The height of each letter represents its information content. **e,** A heatmap of genic mCAS patterns normalized to the flanking region. Each gene was assigned to one of twenty clusters, which is indicated by the number and tick marks on the y-axis. The tick marks on the x-axis indicate the upstream, transcription start, transcription end, and downstream segments of each gene. The boxes around various patterns highlight regions referenced in the main text. **f,** Bar plot of the ratio of the genome-wide mCAC to mCAG in various samples.

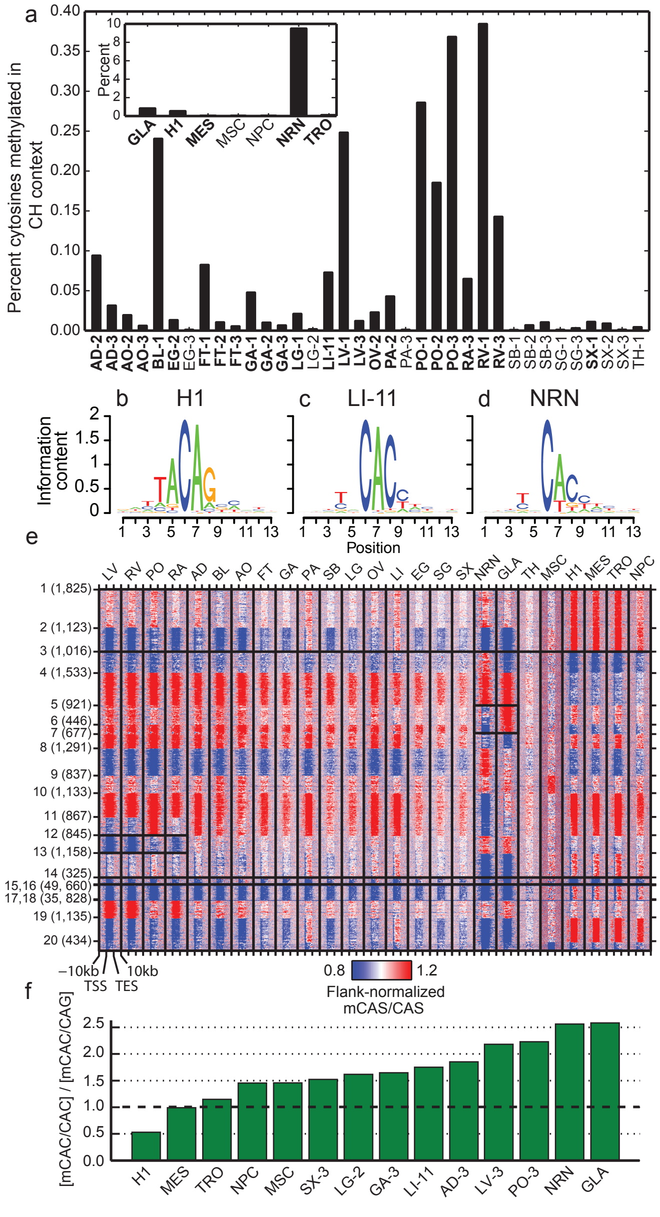

a
Percent cytosines methylated in CH context
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Percent
10
8
6
4
2
0
GLA
H1
MES
MSC
NPC
NRN
TRO
AD-2
AD-3
AO-2
AO-3
BL-1
EG-2
EG-3
FT-1
FT-2
FT-3
GA-1
GA-2
GA-3
LG-1
LG-2
LI-11
LV-1
LV-3
OV-2
PA-2
PA-3
PO-1
PO-2
PO-3
RA-3
RV-1
RV-3
SB-1
SB-2
SB-3
SG-1
SG-3
SX-1
SX-2
SX-3
TH-1
b
H1
c
LI-11
d
NRN
Information content
2
1.5
1
0.5
0
1
3
5
7
9
11
13
Position
e
LV
RV
PO
RA
AD
BL
AO
FT
GA
PA
SB
LG
OV
LI
EG
SG
SX
NRN
GLA
TH
MSC
H1
MES
TRO
NPC
1 (1,825)
2 (1,123)
3 (1,016)
4 (1,533)
5 (921)
6 (446)
7 (677)
8 (1,291)
9 (837)
10 (1,133)
11 (867)
12 (845)
13 (1,158)
14 (325)
15,16 (49, 660)
17,18 (35, 828)
19 (1,135)
20 (434)
−10kb
TSS
TES
10kb
0.8
1.2
Flank-normalized
mCAS/CAS
f
[mCAC/CAC] / [mCAC/CAG]
2.5
2.0
1.5
1.0
0.5
0.0
H1
MES
TRO
NPC
MSC
SX-3
LG-2
GA-3
LI-11
AD-3
LV-3
PO-3
NRN
GLA

Figure 2.4: **Allele-specific Methylation and Expression. a,** Browser screenshot of the increase in female mCH for a gene known to escape X chromosome inactivation (MED14). Sample names are colored by gender (male, black; female, red). **b,** Ratio of mCH level in female vs. male samples across genes with a significant difference in at least one sample. Cells boxed in black denote samples with a statistically significant difference between females and males. **c,** The number of ASM and ASE sites across the triplicated tissues. The top row depicts ASM events (left) and ASE events (right) which are allele-specific in all tissues (black), are variable across tissues (white), or do not possess enough data to tell (grey). The bottom row depicts the distribution of variable sites from the top row that vary by individual (white), tissue (black), or neither (grey).

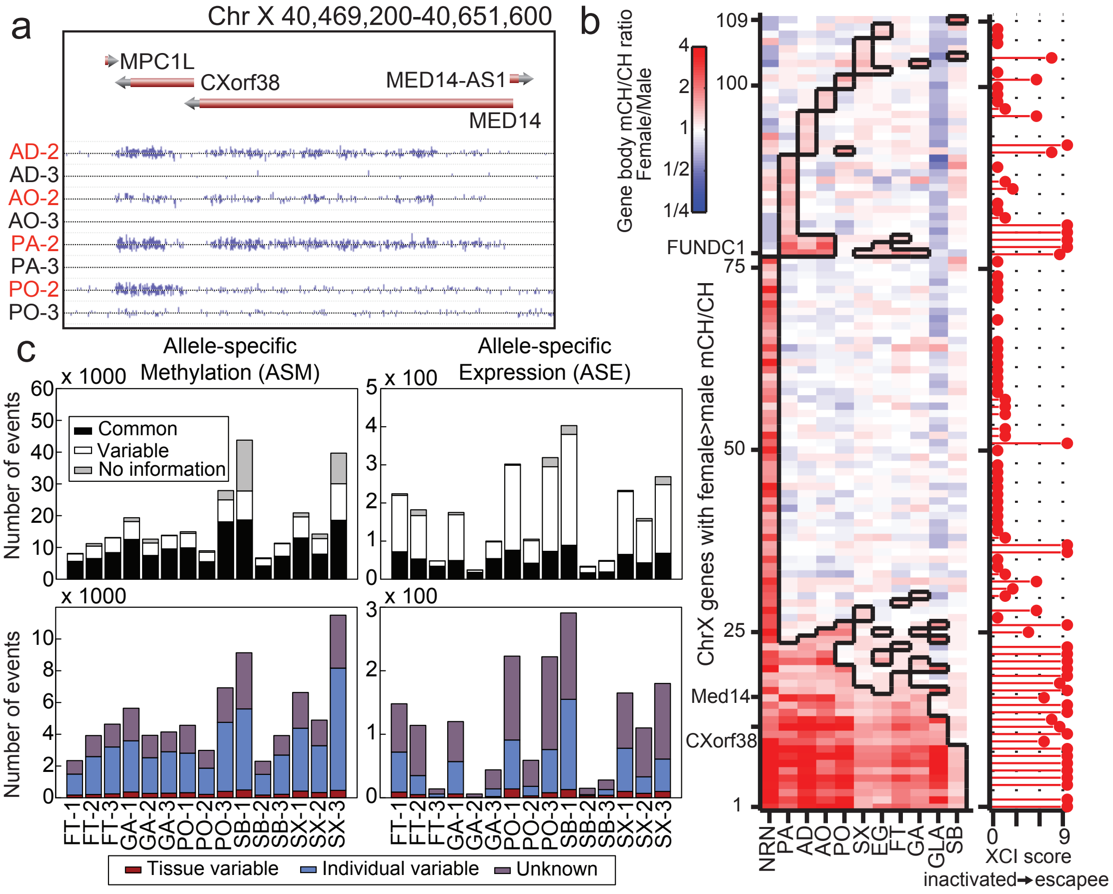

a
Chr X 40,469,200-40,651,600
MPC1L
CXorf38
MED14-AS1
MED14
AD-2
AD-3
AO-2
AO-3
PA-2
PA-3
PO-2
PO-3
b
Gene body mCH/CH ratio
Female/Male
4
2
1
1/2
1/4
109
100
FUNDC1
75
50
25
Med14
CXorf38
1
ChrX genes with female>male mCH/CH
NRN
PA
AD
AO
PO
SX
EG
FT
GA
GLA
SB
0
9
XCI score
inactivated→escapee
c
Allele-specific
Methylation (ASM)
Allele-specific
Expression (ASE)
x 1000
x 100
Number of events
Common
Variable
No information
60
50
40
30
20
10
0
5
4
3
2
1
0
x 1000
x 100
10
8
6
4
2
0
3
2
1
0
FT-1
FT-2
FT-3
GA-1
GA-2
GA-3
PO-1
PO-2
PO-3
SB-1
SB-2
SB-3
SX-1
SX-2
SX-3
Tissue variable
Individual variable
Unknown

Figure 2.5: **Identification of differentially methylation regions (DMRs) and Multidimensional Scaling Analysis. a,** Line plot showing the fraction of differentially methylated CG sites (DMSs, dynamic CGs) out of all CG sites under various methylation difference cutoffs. The methylation difference of a CG site is defined in Ziller et al.[1] **b,** A plot of the first two principal components from the methylation level multi-dimensional scaling. Tissues are shaded by the organ group they belong to as in Figure 1c and 1d. **c-d,** Bar charts of the cumulative amount of variance explained by the first N principal components from the multi-dimensional scaling performed on the methylation levels of all DMRs (c) and the expression levels of all differentially expressed genes (d). **e,** A representative example of enriched GO biological process terms based on the most hypomethylated DMRs from LV-1. **f,** A representative example of enriched mouse phenotype terms based on the most hypomethylated DMRs from LV-1.

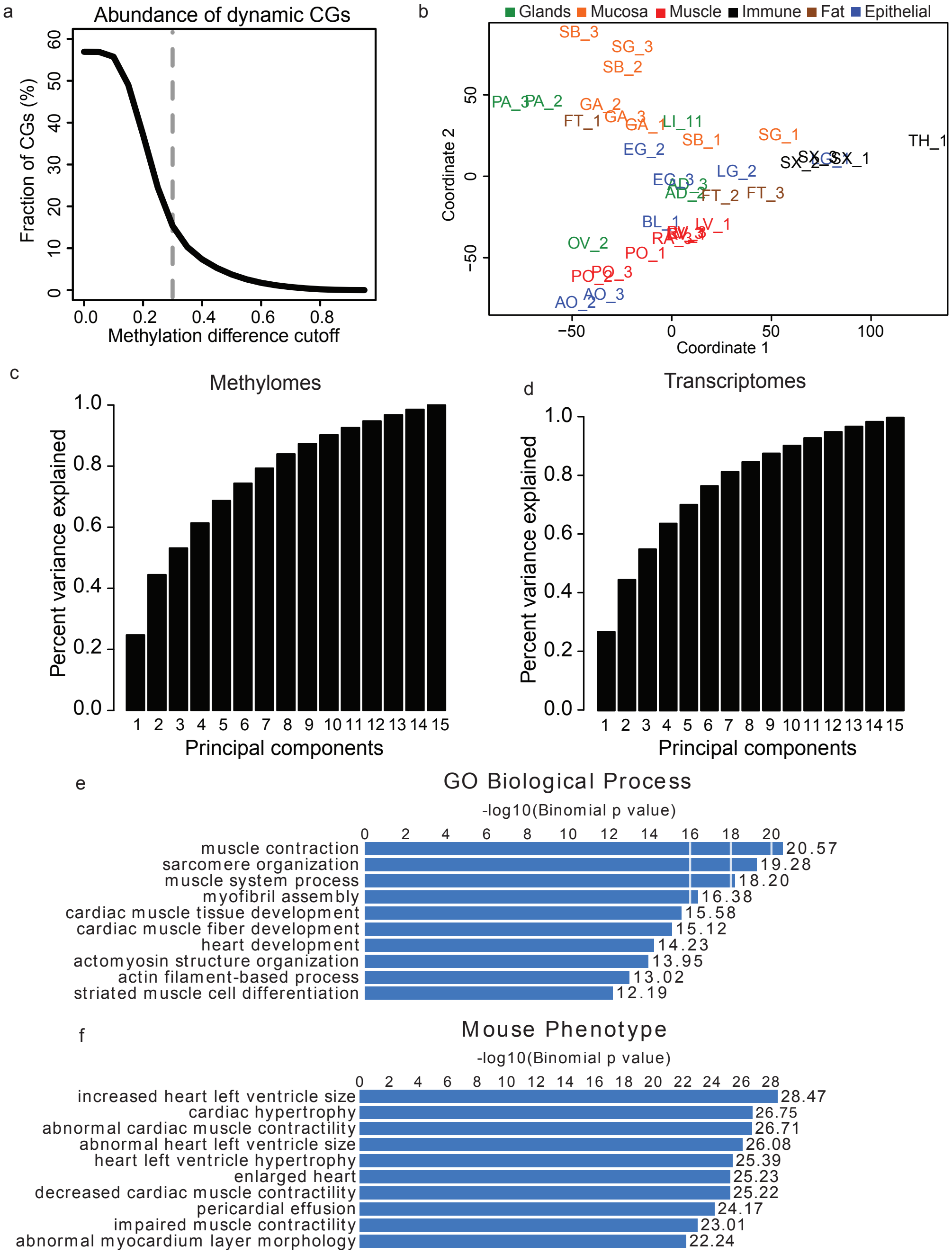

a
Abundance of dynamic CGs
Fraction of CGs (%)
0
10
20
30
40
50
60
0.0
0.2
0.4
0.6
0.8
Methylation difference cutoff
b
🟩 Glands 🟧 Mucosa 🟥 Muscle ⬛ Immune 🟫 Fat 🟦 Epithelial
SB_3
SG_3
SB_2
PA_3
PA_2
GA_2
FT_1
GA_3
GA_1
LI_11
SB_1
SG_1
EG_2
SX_2
SX_3
LG_1
SX_1
TH_1
EG_3
AD_3
LG_2
AD_2
FT_2
FT_3
BL_1
LV_1
RV_3
RA_3
OV_2
PO_1
PO_2
PO_3
AO_2
AO_3
Coordinate 2
50
0
-50
−50
0
50
100
Coordinate 1
c
Methylomes
Percent variance explained
1.0
0.8
0.6
0.4
0.2
0.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Principal components
d
Transcriptomes
Percent variance explained
1.0
0.8
0.6
0.4
0.2
0.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Principal components
e
GO Biological Process
-log10(Binomial p value)
0 2 4 6 8 10 12 14 16 18 20
muscle contraction 20.57
sarcomere organization 19.28
muscle system process 18.20
myofibril assembly 16.38
cardiac muscle tissue development 15.58
cardiac muscle fiber development 15.12
heart development 14.23
actomyosin structure organization 13.95
actin filament-based process 13.02
striated muscle cell differentiation 12.19
f
Mouse Phenotype
-log10(Binomial p value)
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28
increased heart left ventricle size 28.47
cardiac hypertrophy 26.75
abnormal cardiac muscle contractility 26.71
abnormal heart left ventricle size 26.08
heart left ventricle hypertrophy 25.39
enlarged heart 25.23
decreased cardiac muscle contractility 25.22
pericardial effusion 24.17
impaired muscle contractility 23.01
abnormal myocardium layer morphology 22.24

Figure 2.6: **DMRs and their correlation with transcription. a,** A browser screenshot of an example DMR downstream of the TSS. **b,** Expression level of the BIN1 gene which contains the DMR in (a). **c,** The percentages of hypomethylated intragenic DMRs in each class of genomic features. **d-h,** Histone modification profiles of five categories of uiDMRs.

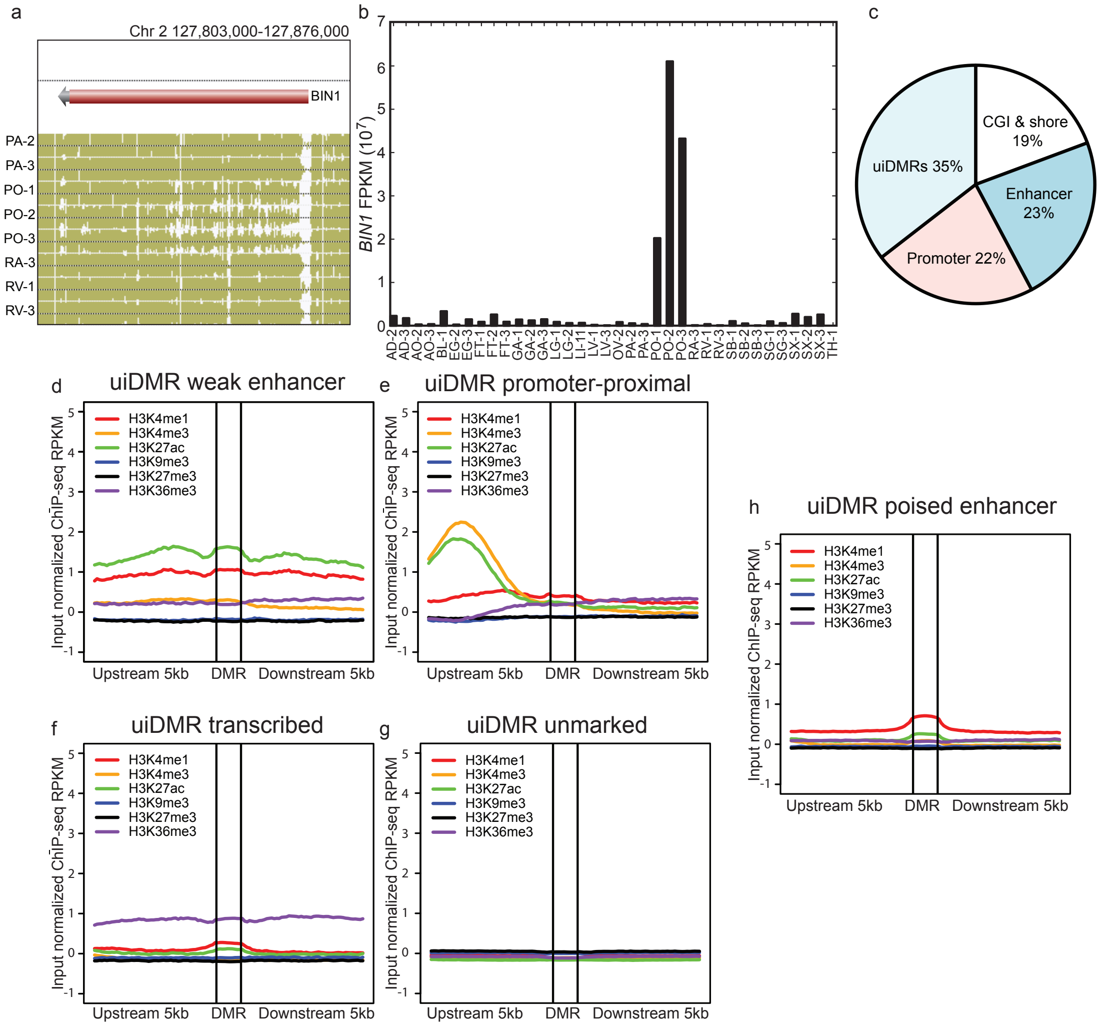

a
Chr 2 127,803,000-127,876,000
BIN1
PA-2
PA-3
PO-1
PO-2
PO-3
RA-3
RV-1
RV-3
b
BIN1 FPKM (10^7)
0
1
2
3
4
5
6
7
AD-2
AD-3
AO-2
AO-3
BL-1
EG-2
EG-3
FT-1
FT-2
FT-3
GA-1
GA-2
GA-3
LG-1
LG-2
LI-11
LV-1
LV-3
OV-2
PA-2
PA-3
PO-1
PO-2
PO-3
RA-3
RV-1
RV-3
SB-1
SB-2
SB-3
SG-1
SG-3
SX-1
SX-2
SX-3
TH-1
c
CGI & shore 19%
Enhancer 23%
Promoter 22%
uiDMRs 35%
d
uiDMR weak enhancer
e
uiDMR promoter-proximal
f
uiDMR transcribed
g
uiDMR unmarked
h
uiDMR poised enhancer
H3K4me1
H3K4me3
H3K27ac
H3K9me3
H3K27me3
H3K36me3
Input normalized ChIP-seq RPKM
-1
0
1
2
3
4
5
Upstream 5kb
DMR
Downstream 5kb

Figure 2.7: **Classification of uiDMR histone profiles and uiDMR properties. a,** heatmap of the histone modification profiles for the five types of uiDMRs. The profiles were plotted for each mark across the DMR and the 5kb upstream and downstream and the colors of each cell indicate the input normalized ChIP-seq RPKM. The colors on the left indicate the group of each profile assigned by k-means clustering (red, weak enhancer; orange, promoter-proximal; green, transcribed; blue, unmarked; black poised enhancer). **b,** A pie chart of the distribution of uiDMRs across the classes defined by k-means clustering.

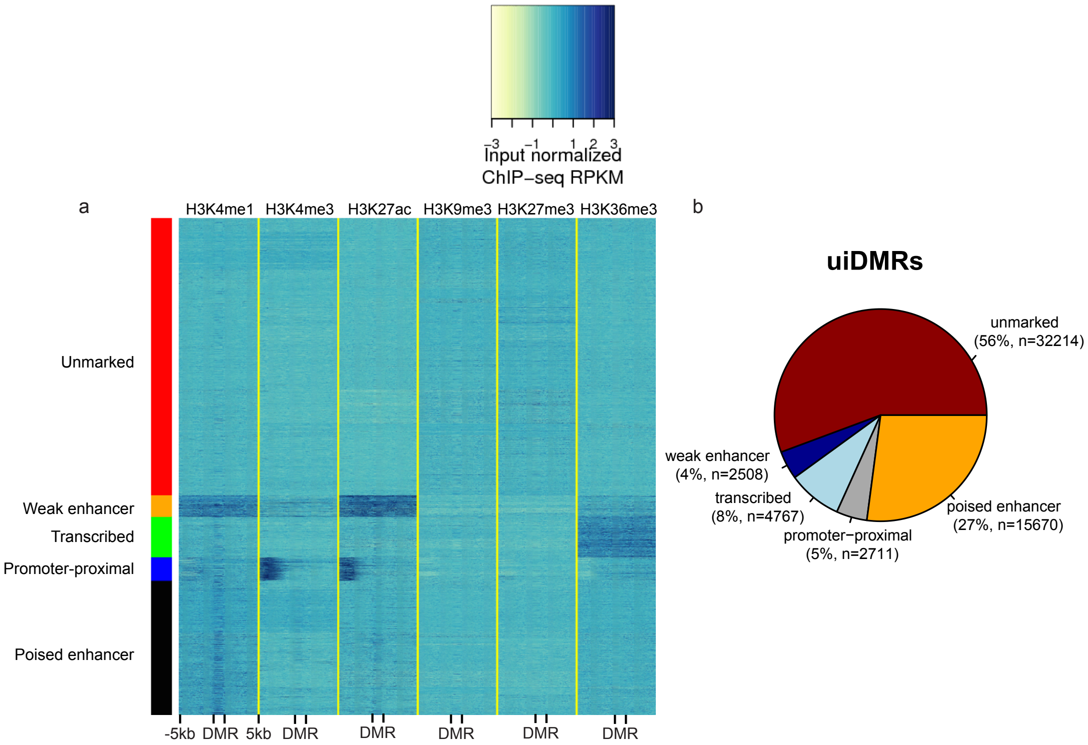

−3 −1 1 2 3
Input normalized
ChIP−seq RPKM
a
H3K4me1
H3K4me3
H3K27ac
H3K9me3
H3K27me3
H3K36me3
Unmarked
Weak enhancer
Transcribed
Promoter-proximal
Poised enhancer
-5kb
DMR
5kb
DMR
DMR
DMR
DMR
DMR
b
uiDMRs
unmarked
(56%, n=32214)
weak enhancer
(4%, n=2508)
transcribed
(8%, n=4767)
promoter−proximal
(5%, n=2711)
poised enhancer
(27%, n=15670)

Figure 2.8: **Classification of promoter histone profiles. a,** A heatmap of the histone modification profiles across strong (rows labeled with red) and unmarked (rows labeled with orange) promoters. The profiles were plotted for each mark across the promoter and the 5kb upstream and downstream and the colors of each cell indicate the input normalized ChIP-seq RPKM. **b-c,** The aggregate profiles for strong and unmarked promoters (b) and (c), respectively. d, The distribution of the Spearman correlation coefficients between the methylation level of different types of hypomethylated intragenic DMRs and the expression of the nearest gene. Notches indicate a confidence interval estimated from 1,000 bootstrap samples.

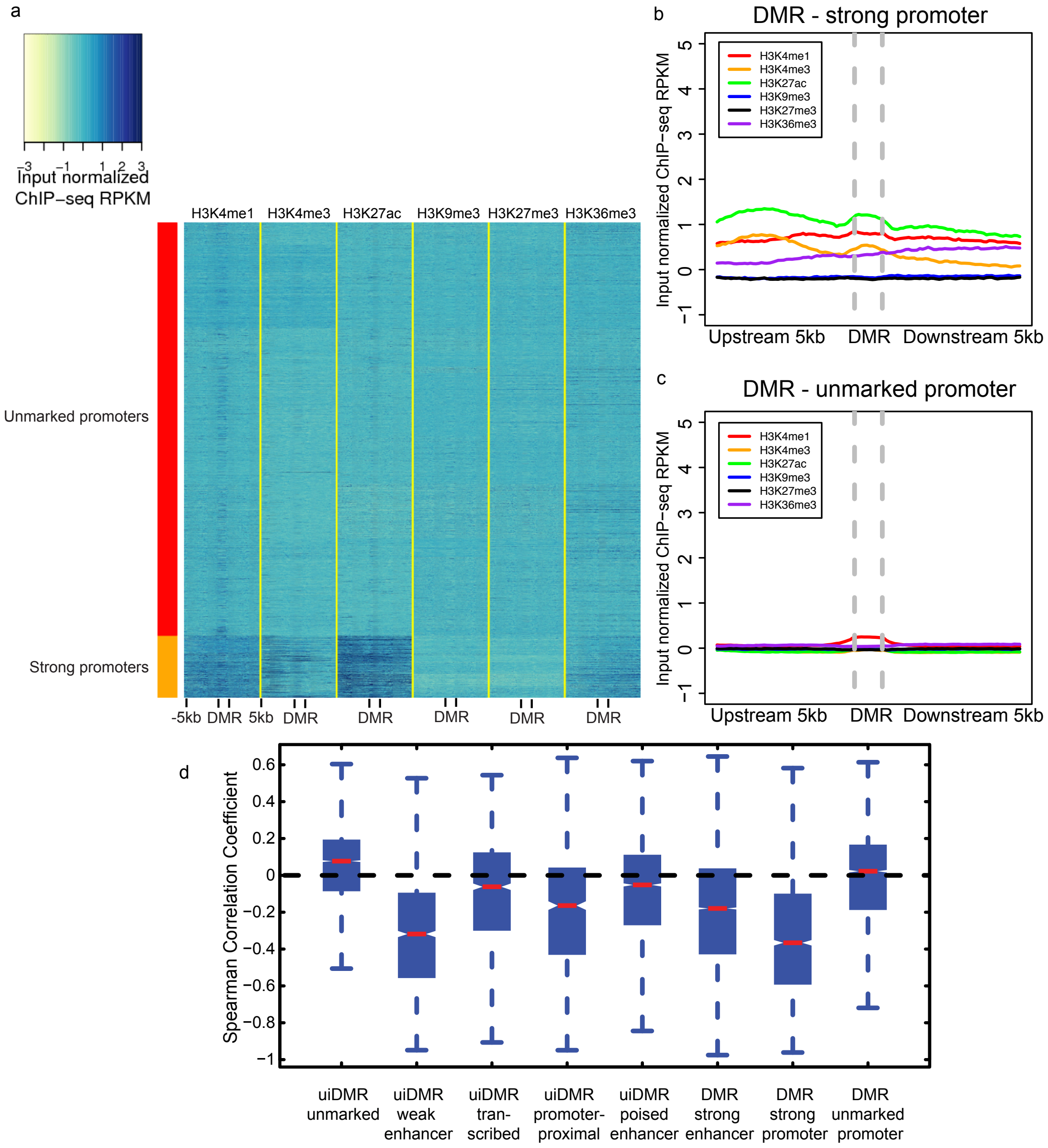

a
Input normalized ChIP−seq RPKM
−3 −1 1 2 3
H3K4me1
H3K4me3
H3K27ac
H3K9me3
H3K27me3
H3K36me3
Unmarked promoters
Strong promoters
-5kb DMR 5kb
DMR
DMR
DMR
DMR
DMR
b
DMR - strong promoter
c
DMR - unmarked promoter
Input normalized ChIP−seq RPKM
−1 0 1 2 3 4 5
H3K4me1
H3K4me3
H3K27ac
H3K9me3
H3K27me3
H3K36me3
Upstream 5kb
DMR
Downstream 5kb
d
Spearman Correlation Coefficient
0.6 0.4 0.2 0 −0.2 −0.4 −0.6 −0.8 −1
uiDMR unmarked
uiDMR weak enhancer
uiDMR tran-scribed
uiDMR promoter-proximal
uiDMR poised enhancer
DMR strong enhancer
DMR strong promoter
DMR unmarked promoter

Figure 2.9: **uiDMR fetal DNase I profiles.** DNase I profiles of various fetal tissues corresponding to the tissues presented in this study. The samples are arranged columnwise by age, and row-wise by fetal tissue. The uiDMR unmarked line represents the DNase I profile of uiDMRs without histone modifications. The DMR enhancer line represents the DNase I profile of DMRs that overlapped an enhancer in a matched tissue in this study (indicated in the row label in parentheses). The shuffled line represents the DNase I profile of uiDMRs randomly shuffled across the genome.

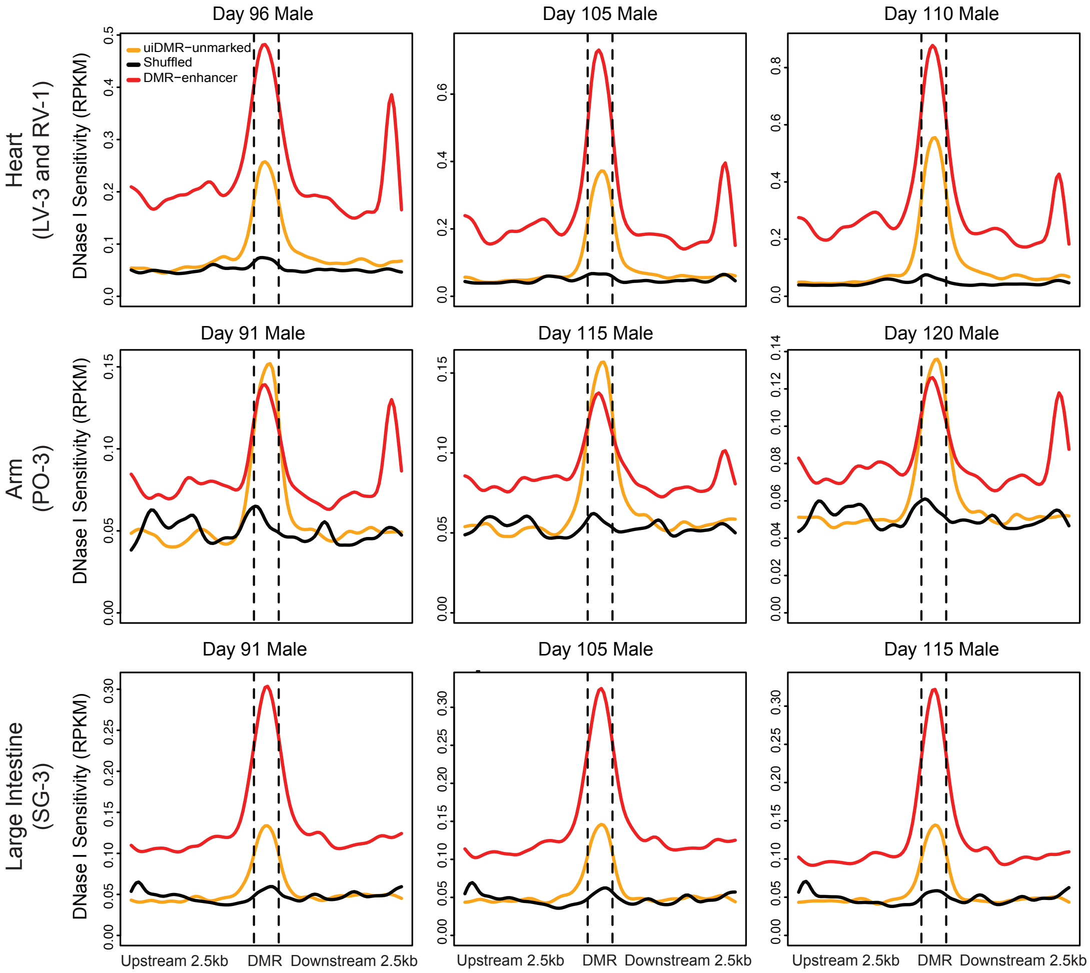

Day 96 Male
Day 105 Male
Day 110 Male
Day 91 Male
Day 115 Male
Day 120 Male
Day 91 Male
Day 105 Male
Day 115 Male
uiDMR−unmarked
Shuffled
DMR−enhancer
Heart
(LV-3 and RV-1)
Arm
(PO-3)
Large Intestine
(SG-3)
DNase I Sensitivity (RPKM)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.8
0.00
0.05
0.10
0.15
0.02
0.04
0.06
0.08
0.12
0.14
0.20
0.25
0.30
Upstream 2.5kb
DMR
Downstream 2.5kb

Figure 2.10: **PMD Features. a,** A browser screenshot (see Figure 1 for description) of an example PMD found in IMR90, PLA, PA-2, and PA-3. RV-1 is included as a representative sample without PMDs. **b,** The distribution of sizes of PMDs in various samples. **c,** A heatmap representation of the overlap between various sets of PMDs. The denominator of the fraction of overlap is determined by the sample on the y-axis. **d-e,** ChIP-seq profiles of the PMD regions defined in PA-2 (c) and IMR90 (d) after shuffling.

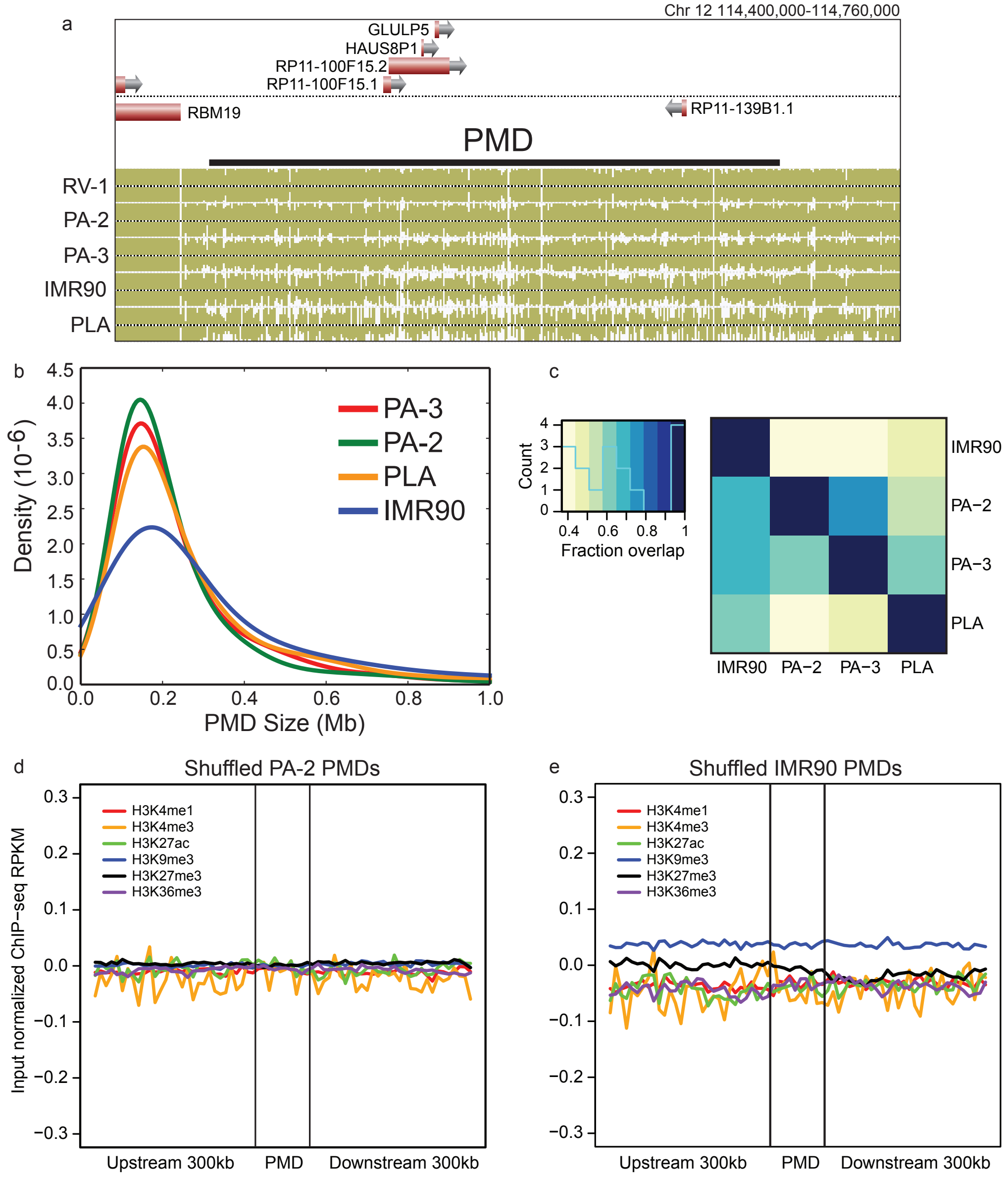

a
Chr 12 114,400,000-114,760,000
GLULP5
HAUS8P1
RP11-100F15.2
RP11-100F15.1
RBM19
RP11-139B1.1
PMD
RV-1
PA-2
PA-3
IMR90
PLA
b
Density (10^-6)
PMD Size (Mb)
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0
PA-3
PA-2
PLA
IMR90
c
Count
4
3
2
1
0
0.4
0.6
0.8
1
Fraction overlap
IMR90
PA−2
PA−3
PLA
IMR90
PA−2
PA−3
PLA
d
Shuffled PA-2 PMDs
Input normalized ChIP−seq RPKM
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
H3K4me1
H3K4me3
H3K27ac
H3K9me3
H3K27me3
H3K36me3
Upstream 300kb
PMD
Downstream 300kb
e
Shuffled IMR90 PMDs
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
H3K4me1
H3K4me3
H3K27ac
H3K9me3
H3K27me3
H3K36me3
Upstream 300kb
PMD
Downstream 300kb

Figure 2.11: **DNMT expression across tissues. a-d,** Bar plots of the expression (measured in log10 FPKMs) of DNMT1 (a), DNMT3A (b), DNMT3B (c), and DNMT3L (d) across various samples.

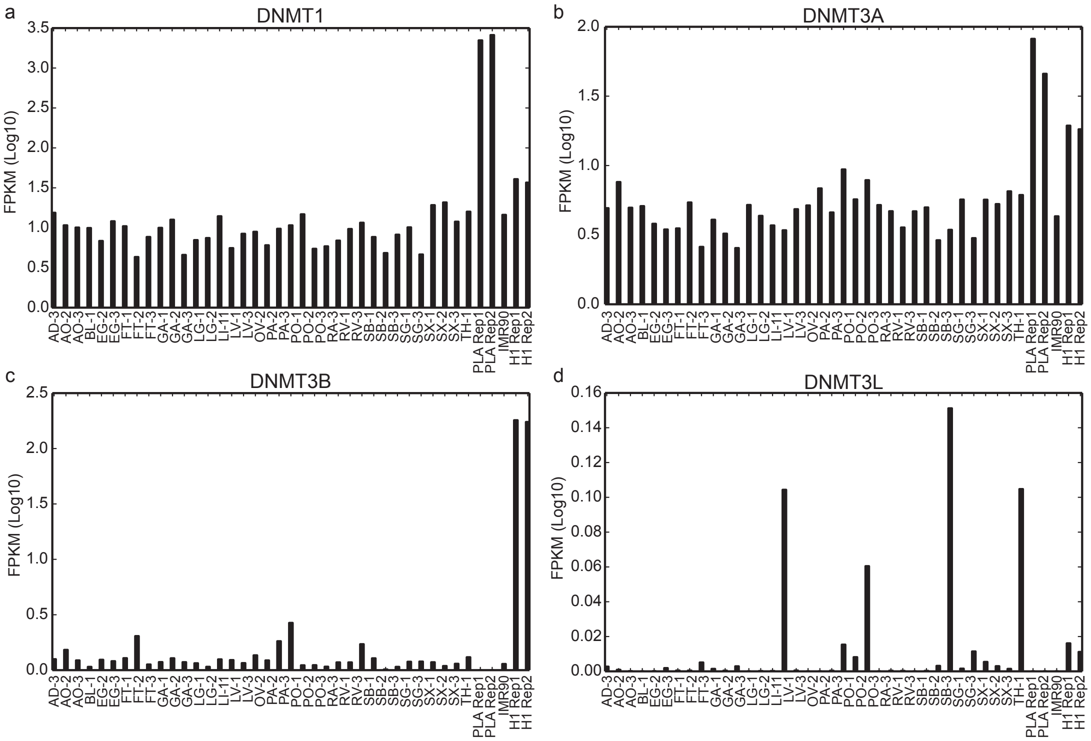

a
DNMT1
FPKM (Log10)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
b
DNMT3A
FPKM (Log10)
0.0
0.5
1.0
1.5
2.0
c
DNMT3B
FPKM (Log10)
0.0
0.5
1.0
1.5
2.0
2.5
d
DNMT3L
FPKM (Log10)
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
AD-3
AO-2
AO-3
BL-1
EG-2
EG-3
FT-1
FT-2
FT-3
GA-1
GA-2
GA-3
LG-1
LG-2
LI-11
LV-1
LV-3
OV-2
PA-2
PA-3
PO-1
PO-2
PO-3
RA-3
RV-1
RV-3
SB-1
SB-2
SB-3
SG-1
SG-3
SX-1
SX-2
SX-3
TH-1
PLA Rep1
PLA Rep2
IMR90
H1 Rep1
H1 Rep2

Figure 2.12: **mCH distribution and correlation. a,** A browser screenshot (see Figure 1 for description) of an example region with non-CG methylation (mCH). Purple and pink ticks are methylated CHG and CHH sites, respectively (H = A, C, or T). Ticks on the forward strand are projected upward from the dotted line and ticks on the reverse strand are projected downward. **b,** The distribution of methylation levels at mCH sites across all samples with a discernible TNCAC motif. Only mCH sites with at least 10 reads and a significant amount of methylation were considered. **c,** Boxplots of the expression values across different quantiles of CAC gene body methylation (Gene body mCAC). **d,** Scatterplot of mCAG vs. mCAC inside gene bodies. **e,** Bar plot of the correlation of mCAG and mCAC inside gene bodies (blue) and the theoretical maximal correlation (red) if mCAC and mCAG are perfectly correlated. **f-h,** The methylation levels of C (upper panel), CG (middle panel) and CH (lower panel) across the read positions for PO-2 (red line) and EG-3 (blue line). Vertical lines indicate the position (10th base from the beginning) where trimming was applied. **i,** mCH motif from PO-2 with the first 10 bases of each read trimmed. j, mCH motif from PO-2 without trimming. k, mCH motif from EG-3 with the first 10 bases of each read trimmed l, mCH motif from EG-3 without trimming. The height of each letter represents its information content (i.e., prevalence).

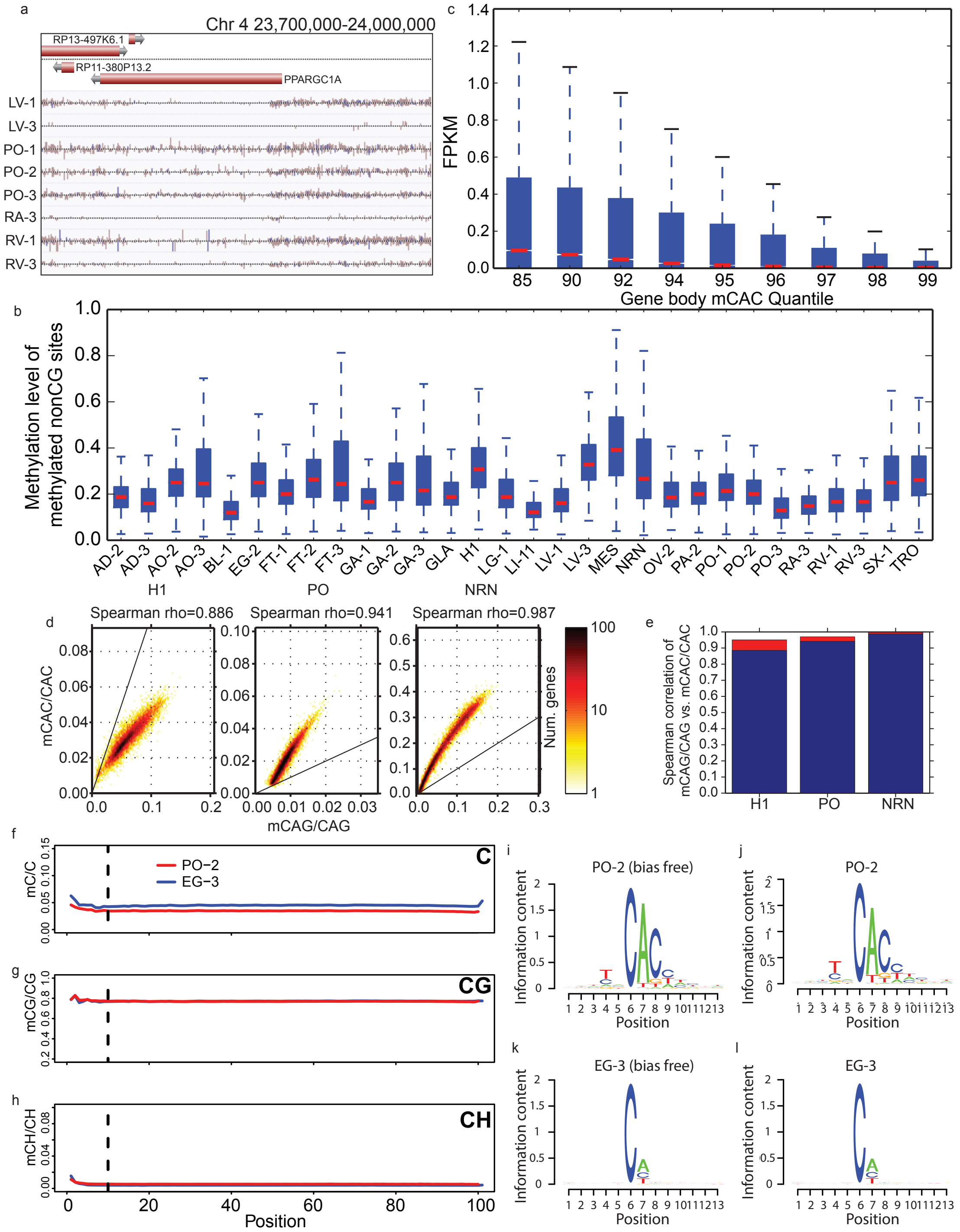

a
Chr 4 23,700,000-24,000,000
RP13-497K6.1
RP11-380P13.2
PPARGC1A
LV-1
LV-3
PO-1
PO-2
PO-3
RA-3
RV-1
RV-3
c
FPKM
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
85
90
92
94
95
96
97
98
99
Gene body mCAC Quantile
b
Methylation level of methylated nonCG sites
1.0
0.8
0.6
0.4
0.2
0.0
AD-2
AD-3
AO-2
AO-3
BL-1
EG-2
FT-1
FT-2
FT-3
GA-1
GA-2
GA-3
GLA
H1
LG-1
LI-11
LV-1
LV-3
MES
NRN
OV-2
PA-2
PO-1
PO-2
PO-3
RA-3
RV-1
RV-3
SX-1
TRO
d
H1
PO
NRN
Spearman rho=0.886
Spearman rho=0.941
Spearman rho=0.987
mCAC/CAC
mCAG/CAG
Num. genes
100
10
1
e
Spearman correlation of mCAG/CAG vs. mCAC/CAC
H1
PO
NRN
f
mC/C
C
PO−2
EG−3
g
mCG/CG
CG
h
mCH/CH
CH
Position
0
20
40
60
80
100
i
PO-2 (bias free)
j
PO-2
k
EG-3 (bias free)
l
EG-3
Information content
Position
1 2 3 4 5 6 7 8 9 10 11 12 13

Figure 2.13: **X chromosome inactivation. a,** Distributions of promoter CG methylation (mCG) levels (mCG/CG), gene body non-CG methylation (mCH) levels (mCH/CH), gene body mCG levels and promoter mCH levels in genes previously reported to express from only one allele (inactivated) or biallelically (escapee). Black ticks show median, and bars indicate 25-75th percentile range. Genes more prone to escaping inactivation have lower promoter mCG, higher gene body mCH, higher gene body mCG and higher promoter mCH in females. **b-e,** Discriminability analysis using **b,** gender-specific gene-body mCH, **c,** promoter mCG, **d,** promoter mCH and **e,** gene body mCG to predict the escapee status of X-linked gene, respectively. Among them, gene body mCH is the most predictive feature of chromosome X inactivation escapees.

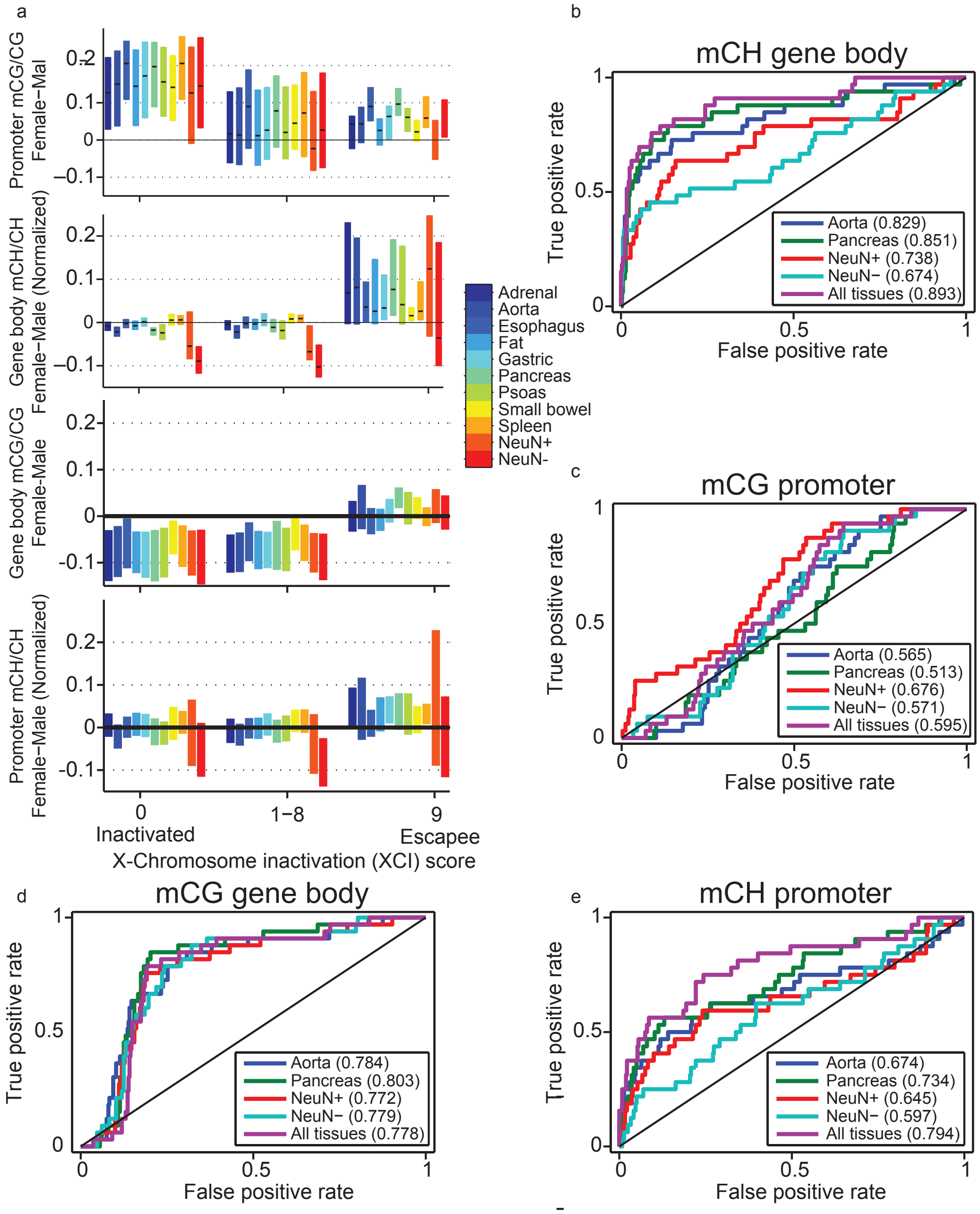
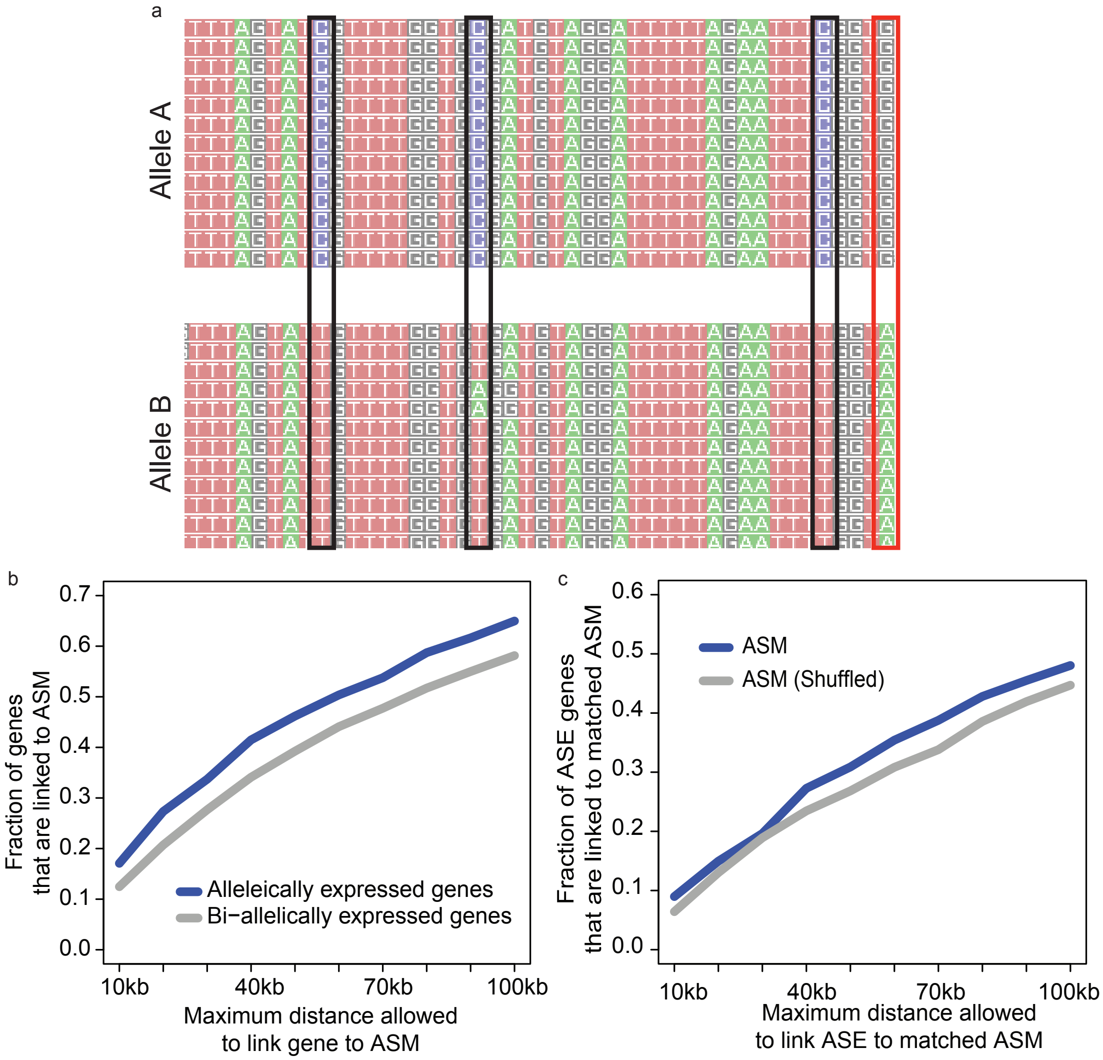

a
Promoter mCG/CG Female−Mal
0.2
0.1
0
−0.1
Gene body mCH/CH Female−Male (Normalized)
0.2
0.1
0
−0.1
Gene body mCG/CG Female-Male
0.2
0.1
0
-0.1
Promoter mCH/CH Female−Male (Normalized)
0.2
0.1
0
−0.1
0
Inactivated
1−8
9
Escapee
X-Chromosome inactivation (XCI) score
Adrenal
Aorta
Esophagus
Fat
Gastric
Pancreas
Psoas
Small bowel
Spleen
NeuN+
NeuN-
b
mCH gene body
True positive rate
False positive rate
0
0.5
1
Aorta (0.829)
Pancreas (0.851)
NeuN+ (0.738)
NeuN− (0.674)
All tissues (0.893)
c
mCG promoter
True positive rate
False positive rate
0
0.5
1
Aorta (0.565)
Pancreas (0.513)
NeuN+ (0.676)
NeuN− (0.571)
All tissues (0.595)
d
mCG gene body
True positive rate
False positive rate
0
0.5
1
Aorta (0.784)
Pancreas (0.803)
NeuN+ (0.772)
NeuN− (0.779)
All tissues (0.778)
e
mCH promoter
True positive rate
False positive rate
0
0.5
1
Aorta (0.674)
Pancreas (0.734)
NeuN+ (0.645)
NeuN− (0.597)
All tissues (0.794)

Figure 2.14: **Allele-specific Methylation and Expression. a,** An example of allele-specific methylation (ASM). Reads that contain a heterozygous SNP (red box) are separated by allele. The number of methylated (reads containing Cs) and unmethylated (reads containing Ts) at adjacent CG sites (black boxes) and tested for differential methylation. **b,** Fraction of allele-specific expressed (ASE) genes (blue) and bi-allelically expressed genes (grey) that have at least one ASM event within a certain distance. Bi-allelically expressed genes were defined as genes that were covered by at least 10 reads and whose p-values given by binomial test for allelic expression were greater than 0.2 (i.e. no significance). **c,** Fraction of ASE genes that were linked to matched ASM event(s) (blue) and matched ASM events with their locations shuffled (grey). b-c are aggregated results using samples from triplicate tissues.


a
Allele A
Allele B
b
Fraction of genes
that are linked to ASM
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Alleleically expressed genes
Bi−allelically expressed genes
10kb
40kb
70kb
100kb
Maximum distance allowed
to link gene to ASM
c
Fraction of ASE genes
that are linked to matched ASM
0.0
0.1
0.2
0.3
0.4
0.5
0.6
ASM
ASM (Shuffled)
10kb
40kb
70kb
100kb
Maximum distance allowed
to link ASE to matched ASM

2.6 References

[1] Michael J. Ziller, Hongcang Gu, Fabian Müller, Julie Donaghey, Linus T-Y Tsai, Oliver Kohlbacher, Philip L. De Jager, Evan D. Rosen, David A. Bennett, Bradley E. Bernstein, Andreas Gnirke, and Alexander Meissner. Charting a dynamic dna methylation landscape of the human genome. *Nature*, 500:477–481, Aug 2013.

[2] Katherine E. Varley, Jason Gertz, Kevin M. Bowling, Stephanie L. Parker, Timothy E. Reddy, Florencia Pauli-Behn, Marie K. Cross, Brian A. Williams, John A. Stamatoyannopoulos, Gregory E. Crawford, Devin M. Absher, Barbara J. Wold, and Richard M. Myers. Dynamic dna methylation across diverse human cell lines and tissues. *Genome Res*, 23:555–567, Mar 2013.

[3] Siddarth Selvaraj, Jesse R Dixon, Vikas Bansal, and Bing Ren. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nature biotechnology*, 31(12):1111–1118, dec 2013. ISSN 1546-1696 (Electronic). doi: 10.1038/nbt.2728.

[4] R. Lister, M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, J. R. Nery, L. Lee, Z. Ye, Q. M. Ngo, L. Edsall, J. Antosiewicz-Bourget, R. Stewart, V. Ruotti, A. H. Millar, J. A. Thomson, B. Ren, and J. R. Ecker. Human dna methylomes at base resolution show widespread epigenomic differences. *Nature*, 462:315–322, 2009.

[5] Rafael A. Irizarry, Christine Ladd-Acosta, Bo Wen, Zhijin Wu, Carolina Montano, Patrick Onyango, Hengmi Cui, Kevin Gabo, Michael Rongione, Maree Webster, Hong Ji, James B. Potash, Sarven Sabunciyan, and Andrew P. Feinberg. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific cpg island shores. *Nat Genet*, 41: 178–186, Feb 2009.

[6] Gary C. Hon, Nisha Rajagopal, Yin Shen, David F. McCleary, Feng Yue, My D. Dang, and Bing Ren. Epigenetic memory at embryonic enhancers identified in dna methylation maps from adult mouse tissues. *Nat Genet*, 45: 1198–1206, Oct 2013.

[7] Denes Hnisz, Brian J. Abraham, Tong Ihn Lee, Ashley Lau, Violaine Saint-André, Alla A. Sigova, Heather A. Hoke, and Richard A. Young. Super-

enhancers in the control of cell identity and disease. *Cell*, 155:934–947, Nov 2013.

[8] Samantha L. Yuen, Ozgur Ogut, and Frank V. Brozovich. Nonmuscle myosin is regulated during smooth muscle contraction. *Am J Physiol Heart Circ Physiol*, 297:H191–H199, Jul 2009.

[9] Matthew D. Schultz, Robert J. Schmitz, and Joseph R. Ecker. 'leveling' the playing field for analyses of single-base resolution dna methylomes. *Trends Genet*, 28:583–585, Dec 2012.

[10] C. Y. McLean, D. Bristor, M. Hiller, S. L. Clarke, B. T. Schaar, C. B. Lowe, A. M. Wenger, and G. Bejerano. Great improves functional interpretation of cis-regulatory regions. *Nat Biotechnol*, 28:495–501, May 2010.

[11] Volker Hovestadt, David T W Jones, Simone Picelli, Wei Wang, Marcel Kool, Paul A Northcott, Marc Sultan, Katharina Stachurski, Marina Ryzhova, Hans-Jorg Warnatz, Meryem Ralser, Sonja Brun, Jens Bunt, Natalie Jager, Kortine Kleinheinz, Serap Erkek, Ursula D Weber, Cynthia C Bartholomae, Christof von Kalle, Chris Lawerenz, Jurgen Eils, Jan Koster, Rogier Versteeg, Till Milde, Olaf Witt, Sabine Schmidt, Stephan Wolf, Torsten Pietsch, Stefan Rutkowski, Wolfram Scheurlen, Michael D Taylor, Benedikt Brors, Jorg Felsberg, Guido Reifenberger, Arndt Borkhardt, Hans Lehrach, Robert J Wechsler-Reya, Roland Eils, Marie-Laure Yaspo, Pablo Landgraf, Andrey Korshunov, Marc Zapatka, Bernhard Radlwimmer, Stefan M Pfister, and Peter Lichter. Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing. *Nature*, 510(7506):537–541, jun 2014. ISSN 0028-0836. URL http://dx.doi.org/10.1038/nature13268http://10.0.4.14/nature13268http://www.nature.com/nature/journal/v510/n7506/abs/nature13268.html{#}supplementary-information.

[12] Alika K. Maunakea, Raman P. Nagarajan, Mikhail Bilenky, Tracy J. Ballinger, Cletus D'Souza, Shaun D. Fouse, Brett E. Johnson, Chibo Hong, Cydney Nielsen, Yongjun Zhao, Gustavo Turecki, Allen Delaney, Richard Varhol, Nina Thiessen, Ksenya Shchors, Vivi M. Heine, David H. Rowitch, Xiaoyun Xing, Chris Fiore, Maximiliaan Schillebeeckx, Steven J. M. Jones, David Haussler, Marco A. Marra, Martin Hirst, Ting Wang, and Joseph F. Costello. Conserved role of intragenic dna methylation in regulating alternative promoters. *Nature*, 466:253–257, Jul 2010.

[13] Akiko Doi, In-Hyun Park, Bo Wen, Peter Murakami, Martin J. Aryee, Rafael Irizarry, Brian Herb, Christine Ladd-Acosta, Junsung Rho, Sabine Loewer, Justine Miller, Thorsten Schlaeger, George Q. Daley, and Andrew P. Feinberg. Differential methylation of tissue- and cancer-specific cpg island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nat Genet*, 41:1350–1353, Dec 2009.

[14] Aimée M. Deaton, Shaun Webb, Alastair R. W. Kerr, Robert S. Illingworth, Jacky Guy, Robert Andrews, and Adrian Bird. Cell type-specific dna methylation at intragenic cpg islands in the immune system. *Genome Res*, 21: 1074–1086, Jul 2011.

[15] Danny Leung, Inkyung Jung, Nisha Rajagopal, Anthony Schmitt, Siddarth Selvaraj, Ah Young Lee, Chia-An Yen, Shin Lin, Yiing Lin, Yunjiang Qiu, Wei Xie, Feng Yue, Manoj Hariharan, Pradipta Ray, Samantha Kuan, Lee Edsall, Hongbo Yang, Neil C Chi, Michael Q Zhang, Joseph R Ecker, and Bing Ren. Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature*, 518(7539):350–354, 2015. ISSN 0028-0836. doi: 10.1038/nature14217. URL http://www.nature.com/doifinder/10.1038/nature14217.

[16] Fereshteh Parviz, Christine Matullo, Wendy D. Garrison, Laura Savatski, John W. Adamson, Gang Ning, Klaus H. Kaestner, Jennifer M. Rossi, Kenneth S. Zaret, and Stephen A. Duncan. Hepatocyte nuclear factor 4alpha controls the development of a hepatic epithelium and liver morphogenesis. *Nat Genet*, 34:292–296, Jul 2003.

[17] Matthew T. Maurano, Richard Humbert, Eric Rynes, Robert E. Thurman, Eric Haugen, Hao Wang, Alex P. Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, Anthony Shafer, Fidencio Neri, Kristen Lee, Tanya Kutyavin, Sandra Stehling-Sun, Audra K. Johnson, Theresa K. Canfield, Erika Giste, Morgan Diegel, Daniel Bates, R. Scott Hansen, Shane Neph, Peter J. Sabo, Shelly Heimfeld, Antony Raubitschek, Steven Ziegler, Chris Cotsapas, Nona Sotoodehnia, Ian Glass, Shamil R. Sunyaev, Rajinder Kaul, and John A. Stamatoyannopoulos. Systematic localization of common disease-associated variation in regulatory dna. *Science*, 337:1190–1195, Sep 2012.

[18] Maria Gutierrez-Arcelus, Tuuli Lappalainen, Stephen B. Montgomery, Alfonso Buil, Halit Ongen, Alisa Yurovsky, Julien Bryois, Thomas Giger, Luciana Romano, Alexandra Planchon, Emilie Falconnet, Deborah Bielser, Maryline

Gagnebin, Ismael Padioleau, Christelle Borel, Audrey Letourneau, Periklis Makrythanasis, Michel Guipponi, Corinne Gehrig, Stylianos E. Antonarakis, and Emmanouil T. Dermitzakis. Passive and active dna methylation and the interplay with genetic variation in gene regulation. *Elife*, 2:e00523, Jan 2013.

[19] Yun Liu, Martin J. Aryee, Leonid Padyukov, M. Daniele Fallin, Espen Hesselberg, Arni Runarsson, Lovisa Reinius, Nathalie Acevedo, Margaret Taub, Marcus Ronninger, Klementy Shchetynsky, Annika Scheynius, Juha Kere, Lars Alfredsson, Lars Klareskog, Tomas J. Ekström, and Andrew P. Feinberg. Epigenome-wide association data implicate dna methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat Biotechnol*, 31:142–147, Feb 2013.

[20] Ryan Lister, Eran A. Mukamel, Joseph R. Nery, Mark Urich, Clare A. Puddifoot, Nicholas D. Johnson, Jacinta Lucero, Yun Huang, Andrew J. Dwork, Matthew D. Schultz, Miao Yu, Julian Tonti-Filippini, Holger Heyn, Shijun Hu, Joseph C. Wu, Anjana Rao, Manel Esteller, Chuan He, Fatemeh G. Haghighi, Terrence J. Sejnowski, M. Margarita Behrens, and Joseph R. Ecker. Global epigenomic reconfiguration during mammalian brain development. *Science*, Jul 2013.

[21] John W Whitaker, Zhao Chen, and Wei Wang. Predicting the human epigenome from DNA motifs. *Nature methods*, 12(3):265–72, 7 p following 272, mar 2015. ISSN 1548-7105 (Electronic). doi: 10.1038/nmeth.3065.

[22] Kryn Stankunas, Ching Shang, Karen Y. Twu, Shih-Chu Kao, Nancy A. Jenkins, Neal G. Copeland, Mrinmoy Sanyal, Licia Selleri, Michael L. Cleary, and Ching-Pin Chang. Pbx/meis deficiencies demonstrate multigenetic origins of congenital heart disease. *Circ Res*, 103:702–709, Sep 2008.

[23] G. C. Hon, R. D. Hawkins, O. L. Caballero, C. Lo, R. Lister, M. Pelizzola, A. Valsesia, Z. Ye, S. Kuan, L. E. Edsall, A. A. Camargo, B. J. Stevenson, J. R. Ecker, V. Bafna, R. L. Strausberg, A. J. Simpson, and B. Ren. Global dna hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. *Genome Res*, 22:246–258, Feb 2012.

[24] B. P. Berman, D. J. Weisenberger, J. F. Aman, T. Hinoue, Z. Ramjan, Y. Liu, H. Noushmehr, C. P. Lange, Dijk CM van, R. A. Tollenaar, Den Berg D. Van, and P. W. Laird. Regions of focal dna hypermethylation and long-range

hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nat Genet*, 44:40–46, Jan 2012.

[25] Diane I. Schroeder, John D. Blair, Paul Lott, Hung On Ken Yu, Danna Hong, Florence Crary, Paul Ashwood, Cheryl Walker, Ian Korf, Wendy P. Robinson, and Janine M. LaSalle. The human placenta methylome. *Proc Natl Acad Sci U S A*, 110:6037–6042, Apr 2013.

[26] R. Lister, M. Pelizzola, Y. S. Kida, R. D. Hawkins, J. R. Nery, G. Hon, J. Antosiewicz-Bourget, R. O'Malley, R. Castanon, S. Klugman, M. Downes, R. Yu, R. Stewart, B. Ren, J. A. Thomson, R. M. Evans, and J. R. Ecker. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature*, 471:68–73, Mar 2011.

[27] Romain Barrès, Megan E. Osler, Jie Yan, Anna Rune, Tomas Fritz, Kenneth Caidahl, Anna Krook, and Juleen R. Zierath. Non-cpg methylation of the pgc-1alpha promoter through dnmt3b controls mitochondrial density. *Cell Metab*, 10:189–198, Sep 2009.

[28] Wei Xie, Matthew D. Schultz, Ryan Lister, Zhonggang Hou, Nisha Rajagopal, Pradipta Ray, John W. Whitaker, Shulan Tian, R. David Hawkins, Danny Leung, Hongbo Yang, Tao Wang, Ah Young Lee, Scott A. Swanson, Jiuchun Zhang, Yun Zhu, Audrey Kim, Joseph R. Nery, Mark A. Urich, Samantha Kuan, Chia-An Yen, Sarit Klugman, Pengzhi Yu, Kran Suknuntha, Nicholas E. Propson, Huaming Chen, Lee E. Edsall, Ulrich Wagner, Yan Li, Zhen Ye, Ashwinikumar Kulkarni, Zhenyu Xuan, Wen-Yu Chung, Neil C. Chi, Jessica E. Antosiewicz-Bourget, Igor Slukvin, Ron Stewart, Michael Q. Zhang, Wei Wang, James A. Thomson, Joseph R. Ecker, and Bing Ren. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell*, 153:1134–1148, May 2013.

[29] Laura Carrel and Huntington F Willard. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature*, 434(7031):400–404, mar 2005. ISSN 0028-0836. URL http://dx.doi.org/10.1038/nature03479http://www.nature.com/nature/journal/v434/n7031/suppinfo/nature03479{_}S1.html.

[30] Amy J. Wagers and Irving L. Weissman. Plasticity of adult stem cells. *Cell*, 116:639–648, Mar 2004.

[31] Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with bowtie 2. *Nat Methods*, 9:357–359, Apr 2012.

[32] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A. DePristo. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome Res*, 20:1297–1303, Sep 2010.

[33] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17, 2011.

[34] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, 10:R25, Jan 2009.

[35] Tao Wang, Jie Liu, Li Shen, Julian Tonti-Filippini, Yun Zhu, Haiyang Jia, Ryan Lister, Joseph Ecker, A. Harvey Millar, Bing Ren, and Others. STAR: an integrated solution to management and visualization of sequencing data. *Bioinformatics*, page btt558, 2013.

[36] William Perkins, Mark Tygert, and Rachel Ward. An introduction to how chi-square and classical exact tests often wildly misreport significance and how the remedy lies in computers. Jan 2012.

[37] Tim Bancroft, Chuanlong Du, and Dan Nettleton. Estimation of false discovery rate using sequential permutation p-values. *Biometrics*, 69:1–7, Mar 2013.

[38] Kasper D. Hansen, Benjamin Langmead, and Rafael A. Irizarry. Bsmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol*, 13:R83, Oct 2012.

[39] Hao Feng, Karen N Conneely, and Hao Wu. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic acids research*, 42(8):e69, apr 2014. ISSN 1362-4962 (Electronic). doi: 10.1093/nar/gku154.

[40] Deqiang Sun, Yuanxin Xi, Benjamin Rodriguez, Hyun Jung Park, Pan Tong, Mira Meong, Margaret A Goodell, and Wei Li. MOABS: model based analysis of bisulfite sequencing data. *Genome biology*, 15(2):R38, feb 2014. ISSN 1474-760X (Electronic). doi: 10.1186/gb-2014-15-2-r38.

[41] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L. Salzberg. Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*, 14:R36, Apr 2013.

[42] Jennifer Harrow, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L. Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, If Barnes, Alexandra Bignell, Veronika Boychenko, Toby Hunt, Mike Kay, Gaurab Mukherjee, Jeena Rajan, Gloria Despacio-Reyes, Gary Saunders, Charles Steward, Rachel Harte, Michael Lin, Cédric Howald, Andrea Tanzer, Thomas Derrien, Jacqueline Chrast, Nathalie Walters, Suganthi Balasubramanian, Baikang Pei, Michael Tress, Jose Manuel Rodriguez, Iakes Ezkurdia, Jeltje van Baren, Michael Brent, David Haussler, Manolis Kellis, Alfonso Valencia, Alexandre Reymond, Mark Gerstein, Roderic Guigó, and Tim J. Hubbard. Gencode: the reference human genome annotation for the encode project. *Genome Res*, 22:1760–1774, Sep 2012.

[43] Adam Roberts, Harold Pimentel, Cole Trapnell, and Lior Pachter. Identification of novel transcripts in annotated genomes using rna-seq. *Bioinformatics*, 27:2325–2329, Sep 2011.

[44] Steven P. Lund, Dan Nettleton, Davis J. McCarthy, and Gordon K. Smyth. Detecting differential expression in rna-sequence data using quasi-likelihood with shrunken dispersion estimates. *Stat Appl Genet Mol Biol*, 11, Jan 2012.

[45] Davis J. McCarthy, Yunshun Chen, and Gordon K. Smyth. Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. *Nucleic Acids Res*, 40:4288–4297, May 2012.

[46] Gregory R. Warnes, Ben Bolker, Lodewijk Bonebakker, Robert Gentleman, Wolfgang Huber Andy Liaw, Thomas Lumley, Martin Maechler, Arni Magnusson, Steffen Moeller, Marc Schwartz, and Bill Venables. *gplots: Various R Programming Tools for Plotting Data*, 2016. URL http://CRAN.R-project.org/package=gplots. R package version 3.0.1.

[47] Laurence R. Meyer, Ann S. Zweig, Angie S. Hinrichs, Donna Karolchik, Robert M. Kuhn, Matthew Wong, Cricket A. Sloan, Kate R. Rosenbloom, Greg Roe, Brooke Rhead, Brian J. Raney, Andy Pohl, Venkat S. Malladi, Chin H. Li, Brian T. Lee, Katrina Learned, Vanessa Kirkup, Fan Hsu, Steve Heitner, Rachel A. Harte, Maximilian Haeussler, Luvina Guruvadoo, Mary Goldman, Belinda M. Giardine, Pauline A. Fujita, Timothy R. Dreszer, Mark Diekhans, Melissa S. Cline, Hiram Clawson, Galt P. Barber, David Haussler, and W. James Kent. The ucsc genome browser database: extensions and updates 2013. *Nucleic Acids Res*, 41:D64–D69, Jan 2013.

[48] Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C. Lin, Peter Laslo, Jason X. Cheng, Cornelis Murre, Harinder Singh, and Christopher K. Glass. 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.

[49] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*, 33:1, Jan 2010.

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

[51] Elodie Portales-Casamar, Supat Thongjuea, Andrew T. Kwon, David Arenillas, Xiaobei Zhao, Eivind Valen, Dimas Yusuf, Boris Lenhard, Wyeth W. Wasserman, and Albin Sandelin. Jaspar 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res*, 38:D105–D110, Jan 2010.

[52] Kimberly Robasky and Martha L. Bulyk. 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.

[53] Zhi Xie, Shaohui Hu, Seth Blackshaw, Heng Zhu, and Jiang Qian. hpdi: a database of experimental human protein-dna interactions. *Bioinformatics*, 26: 287–289, Jan 2010.

[54] Arttu Jolma, Jian Yan, Thomas Whitington, Jarkko Toivonen, Kazuhiro R. Nitta, Pasi Rastas, Ekaterina Morgunova, Martin Enge, Mikko Taipale, Gonghong Wei, Kimmo Palin, Juan M. Vaquerizas, Renaud Vincentelli, Nicholas M. Luscombe, Timothy R. Hughes, Patrick Lemaire, Esko Ukkonen, Teemu Kivioja, and Jussi Taipale. Dna-binding specificities of human transcription factors. *Cell*, 152:327–339, Jan 2013.

[55] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[56] Andreas Heger, Caleb Webber, Martin Goodson, Chris P. Ponting, and Gerton Lunter. Gat: a simulation framework for testing the association of genomic intervals. *Bioinformatics*, Jun 2013.

[57] Oliver Bembom. *seqLogo: Sequence logos for DNA sequence alignments*. R package version 1.36.0.

[58] Wei-Chun Kao and Yun S Song. naiveBayesCall: an efficient model-based base-calling algorithm for high-throughput sequencing. *Journal of computational biology : a journal of computational molecular cell biology*, 18(3):365–377, mar 2011. ISSN 1557-8666 (Electronic). doi: 10.1089/cmb.2010.0247.

[59] Vikas Bansal and Vineet Bafna. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics (Oxford, England)*, 24(16):i153–9, aug 2008. ISSN 1367-4811 (Electronic). doi: 10.1093/bioinformatics/btn298.

[60] Brian L. Browning and Sharon R. Browning. Improving the accuracy and efficiency of identity by descent detection in population data. *Genetics*, 2013. ISSN 0016-6731. doi: 10.1534/genetics.113.150029. URL http://www.genetics.org/content/early/2013/03/25/genetics.113.150029.

Chapter 3

Improved regulatory element prediction based on tissue-specific local epigenomic signatures

3.1 Abstract

Accurate enhancer identification is critical for understanding the spatiotemporal transcriptional regulation during development as well as the functional impact of disease-related non-coding genetic variants. Computational methods have been developed to predict the genomic locations of active enhancers based on histone modifications but the accuracy and resolution of these methods remain limited. Here, we present a novel algorithm REPTILE, which integrates histone modification and whole genome cytosine DNA methylation profiles to identify the precise location of enhancers. We tested the ability of REPTILE to identify en-

hancers previously validated in reporter assays. Compared to existing methods, REPTILE shows consistently superior performance across diverse cell and tissue types, and the enhancer locations are significantly more refined. We show that by incorporating base-resolution methylation data, REPTILE greatly improves upon current methods for annotation of enhancers across a variety of cell and tissue types. REPTILE is available at https://github.com/yupenghe/REPTILE/.

3.2 Significance Statement

In mammals, when and where a gene is transcribed is primarily regulated by the activity of regulatory DNA elements, or enhancers. Genetic mutation disrupting enhancer function is emerging as one of the major causes of human diseases. However, our knowledge remains limited about the location and activity of enhancers in the numerous and distinct cell types and tissues. Here, we develop a new computational approach, REPTILE, to precisely locate enhancers based on genome-wide DNA methylation and histone modification profiling. We systematically tested REPTILE on a variety of human and mouse cell types and tissues. Compared to existing methods, we found that enhancer predictions from REPTILE are more likely to be active in vivo and the predicted locations are more accurate.

3.3 Introduction

In mammals, genes are transcribed in a temporally and spatially specific manner during development. The precise regulation of gene expression is primarily driven by the activity of distal regulatory sequences, known as enhancers. Disruption of enhancers can cause developmental abnormalities and diseases [1, 2, 3, 4, 5, 6]. Moreover, the vast majority of genetic variants associated with human diseases by Genome-Wide Association Studies (GWAS) lie in noncoding regions, which potentially affect gene transcription and contribute to diseases through disrupting enhancer activity[7, 8]. In order to identify causal noncoding variants and understand their functional consequences, comprehensive and methods for accurate enhancer annotation are essential.

Enhancers are bound by transcription factors (TFs), which in turn recruit co-factors such as the histone acetyltransferase EP300 to achieve transcription activation of target genes from a distance[9]. Active enhancers are generally located in accessible chromatin and marked by enrichment of histone H3 lysine 4 monomethylation (H3K4me1) and H3 lysine 27 acetylation (H3K27ac)[10, 11, 12]. Enrichment of histone modifications in the genome can be determined by chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-seq).

Computational approaches have been developed to predict active enhancers from the combinations of these genome-wide profiles (see review[13] for a list of representative methods). They generally use machine learning algorithms to learn the histone modification profiles of putative enhancers active in a given cell/tissue type and then predict enhancers in additional cell/tissue types. Although they have proven to be useful, these methods have several important limitations. First, the centers and boundaries of enhancer predictions are not well defined because of the broad enrichment of histone modifications in regions around enhancers. Second, existing methods often perform worse when tested on cells and tissues other than the cell/tissue types used for training of the algorithm. Third, existing methods consider only one cell/tissue type at a time, and thus neglect potentially useful information about the variation between cell/tissue types.

To address these limitations, we developed REPTILE (Regulatory Element Prediction based on TIssue-specific Local Epigenetic marks), a novel algorithm to predict enhancers by integrating whole genome, base-resolution cell/tissue-specific DNA methylation data along with histone modification data. Cytosine DNA methylation (mC) is a type of chemical modification that plays critical roles in gene regulation, transposon repression and the determination of cell identity[14, 15, 16, 17]. In mammalian genomes, it occurs in both CG and non-CG contexts[18, 19, 20, 21, 22] and can be quantified at nucleotide resolution using whole-genome bisulfite sequencing (WGBS)[18]. In this study, we consider only the most prevalent form of cytosine methylation (mCG). Transcription factor binding sites (TFBSs) are generally depleted of mCG[18, 23]. Whether mCG affects binding affinity is unclear for the majority of transcription factors (TFs), although recent studies suggest that there can be significant alteration of binding affinity[24, 25, 26]. The anti-correlation of mCG and TF binding is predictive in inferring TFBS[27] and enhancers[23, 28]. These observations led us to take advantage of mCG depletion as a high-resolution (1bp depending on density of CG sites) enhancer signature which is complementary to the lower-resolution histone modification data derived from ChIP-seq experiments (with fragment size ranged from 200bp to 600bp after sonication)[29]. Our results indicate that by incorporating mCG data, REPTILE achieves higher prediction accuracy and produces higher-resolution enhancer predictions than existing methods that rely solely on histone modification profiles.

3.4 Results

3.4.1 The REPTILE algorithm.

We designed REPTILE based on three observations: 1) Active enhancers, which are bound by TFs in certain cells and tissues, show cell/tissue-specific hypomethylated and such anti-correlation is an informative feature in predicting enhancers. It has been shown that regions that are differentially methylated across diverse cell and tissue types (also known as differentially methylated regions, DMRs) strongly overlap with enhancers[19, 20, 30]. 2) With base-resolution mCG data, the centers and boundaries of DMRs can be accurately defined, which may be informative in identifying the precise location of enhancers. 3) The known enhancers[31, 32] (2kb) are generally much larger than TFBSs (10-20bp) and likely include sequences that contribute little to enhancer activity. We used the term query region to describe such large regions where a small fraction of the sequences may have a regulatory role. Query regions also refer to negative regions (that showed no observable enhancer activity) and the genomic windows used by enhancer prediction methods. Since a large portion of an active query region may have little contribution to its enhancer activity, the epigenomic signature of the whole active query region may not be an ideal approximation to the epigenomic state of the bona fide regulatory sequences within it. To address this issue, we used DMRs (500bp) to pinpoint the possible regulatory sub-regions within the query regions and to capture informative local epigenomic signatures in both enhancer model training and prediction generation processes (Figure 3.1A-B).

Specifically, the REPTILE algorithm involves four major steps (Figure 3.1C).

First, DMRs are identified by comparing the mCG profiles of the target sample (in which enhancers will be predicted) and several different cell/tissue types (which serve as reference) (see Methods). Next, REPTILE integrates epigenomic data and represents each DMR or query region as a feature vector, where each element is the value of either the intensity or the intensity deviation of an epigenetic mark (Figure 3.1D). The intensity deviation feature captures the epigenomic variation between cell/tissue types and is a unique aspect of REPTILE, whereas existing methods rely on data of a single cell/tissue type (Figure 3.6A; See Methods). In the third step, REPTILE learns a model of enhancer epigenomic signatures from the feature values of (putative) known enhancers and negative regions as well as the DMRs within them. This model contains two random forest[33] classifiers, which predict enhancer activities of query regions and DMRs based on their own epigenomic signature (see Methods). In the last step, REPTILE uses the learned model to calculate enhancer confidence scores for DMRs and query regions, based on which the final predictions are generated (see Methods).

3.4.2 Training computational models for human and mouse enhancers.

To evaluate the prediction accuracy of REPTILE, we systematically compared REPTILE with four widely used enhancer prediction methods, PEDLA[34], RFECS[35], DELTA[36] and CSIANN[37] using data from a wide variety of human and mouse cells and tissues (Figure 3.6B-D; Methods). These methods all use machine learning techniques to predict active enhancers based on histone modification profiles, while PEDLA also considers evolutionary conservation (Supplemental

Methods). Unless specifically stated, six histone modifications were used in these analyses, including H3K4me1, H3K4me2, H3K4me3, H3K27me3, H3K27ac and H3K9ac (Methods). Notably, REPTILEs utilizes mCG information in addition to histone marks.

For each method, we trained a model (a set of parameters) for human enhancers using epigenomic data from H1 human embryonic stem cells and a model for mouse enhancers using data from mouse embryonic stem cells (mESCs). During the training process, EP300 binding sites were used as putative active enhancers (positive instances), while promoters and randomly chosen genomic regions were used as negative instances (Supplemental Methods). When the REPTILE human enhancer model was trained, data of four H1 derived cell types was also included as a reference and DMRs were called for the methylomes of H1 and these cell types. During training of the REPTILE mouse enhancer model, data for eight mouse tissues from E11.5 embryo was used as the reference and DMRs were called across the methylomes of mESCs and all these tissues. In the prediction step, all samples except the target sample were used as the reference. For example, when we applied REPTILE to generate enhancer predictions for E11.5 forebrain, mESCs and the remaining E11.5 tissues were used as the reference.

Unless explicitly stated, all putative enhancers in human cell types and tissues were generated for each method using the human enhancer model, trained using H1 data as described above. Similarly, all enhancer predictions in mouse cell types and tissues were based on the mouse enhancer model, trained using data from mESCs.

3.4.3 REPTILE shows superior prediction accuracy compared to existing methods.

We first used cross validation to evaluate the learned human enhancer models and mouse enhancer models in H1 and mESCs, where the models were trained. In both cell types, REPTILE showed the best performance among all of the tested methods (Figure 3.7A-B). In addition, we found that in H1 cells, putative enhancers from REPTILE and RFECS had the greatest overlaps with distal TFBSs and/or distal open chromatin regions (DHSs), while REPTILE outperformed all other methods in mESCs (Figure 3.2A-B; Supplemental Methods). Also, REPTILE showed one of the highest validation rates (fraction of predictions that are within 1kb to distal DHSs but not in promoters) and one of the lowest misclassification rates (fraction of predictions that are within promoters; Figure 3.8A-D). We then tested REPTILE on the 211 experimentally validated regions in mESCs from Yue et al.[32] and it showed superior performance compared to all other methods (Figure 3.2C; Supplemental Methods). Furthermore, we found that REPTILE predictions recaptured the most distal regulatory DNA elements that were identified by multiplexed editing regulatory assay (MERA), a high-throughput genome mutation screening approach[38] (Figure 3.7C; Supplemental Methods).

Since training datasets (e.g. EP300 data) are often not available for the cells or tissues of interest (target samples), it is extremely desirable that the enhancer model learned on one cell/tissue also performs well on other cell/tissue types. To assess this, we applied the models trained on human embryonic stem cell (H1) data to four H1 derived human cell lines and the models trained on mESCs to eight tissues from E11.5 mouse embryo. In human cell types, REPTILE and DELTA show the highest validation rate and the lowest misclassification rate compared to other methods, while REPTILE performed the best for mouse enhancer prediction (Figure 3.2D-G; Figure 3.9 and Figure 3.10). REPTILE predictions in E11.5 mouse tissues recapitulated several newly in vivo validated enhancers in E11.5 mouse embryo (Figure 3.2H; Supplemental Table S1 and Supplemental Methods). We then tested REPTILE on in vivo experimentally validated regions and found it achieved the best performance for all test datasets except in E11.5 midbrain and heart where it ranked second (Figure 3.2C). Taken together, theses results demonstrate REPTILEs superior prediction accuracy in both human and mouse cell/tissue types over existing methods, when training and prediction were performed on different samples.

3.4.4 The resolution of REPTILE predictions is better than existing methods.

Next, to measure the resolution of enhancer prediction methods, we calculated the average distance between the center of each prediction and the nearest distal DHS (see Methods). We found a higher percentage (82%) of REPTILE mESCs predictions had distal DHS nearby (within 1kb) compared to all other methods (77%; Figure 3.8E). For H1 cells, its overlap (90%) ranked second, which is only slightly lower than RFECS predictions (91%) (Figure 3.8F). Among these predictions, the centers of RFECS predictions are on average 36bp (H1) and 44bp (mESCs) closer to the nearest distal DHSs than REPTILE predictions, which ranked second (Figure 3.8G-H). The results highlight RFECSs superior prediction resolution in the training cell lines (H1 and mESCs), whereas REPTILEs performance is comparable; both outperformed all other methods.

However, we found that REPTILE achieved much better prediction resolution than all other methods when applied to cell/tissue types different than the training data. In H1 derived human cells, the enhancer predictions made by REPTILE are, on average, over 24bp closer to the nearest distal DHSs compared to other methods, including RFECS (Figure 3.3A). On average, 85% of REPTILE predictions are supported by nearby distal DHSs, which ranked second, only slightly lower than DELTA (86%, Figure 3.3B). In tissues from E11.5 mouse embryo, REPTILE predictions are, on average, over 58bp closer to the nearest distal DHSs than the other methods and 92% of the REPTILE predictions are close to distal open chromatin regions, outperforming all other methods (84%; Figure 3.3C-D).

3.4.5 Identifying the transcription factors functionally related to each cell type using REPTILE enhancers.

Enhancers are frequently bound by TFs that are critical to the function of cells and tissues. In H1 and H1 derived cell lineages, we found that the predicted enhancers from REPTILE and other methods are enriched for the DNA motifs that are bound by the TFs (or complex) known to function in these cell lines (Figure 3.4; Supplemental Methods). Motif analysis of REPTILE enhancers recapitulated the enrichment of TF binding motifs in 25 out of the 27 cases (92.6%). Furthermore, in most cases (21/27, 77.8%), the TF binding motif showed stronger enrichment in REPTILE enhancers than in the putative enhancers from other methods. Notably, in the trophoblast-like cell lineage (TRO), the average enrichment of the TF motifs nearly doubled in enhancers from REPTILE compared to other methods (2.5 fold versus 1.3 fold; Figure 3.4). These results indicate that REPTILE enhancer predictions facilitate the discovery of functionally related TFs in a given cell type by accurately pinpointing the location of their binding motifs.

3.4.6 REPTILE enhancers are enriched for non-coding GWAS SNPs and associated with increased expression of target genes.

Non-coding disease-associated genetic variants are enriched in the regulatory elements of related cell types and tissues[7]. Stronger tissue-specific enrichment of such variants in putative enhancers of related tissues or cell types is likely indicative of better prediction accuracy and resolution. Therefore, we employed enrichment as a metric for the evaluation of enhancer prediction methods.

First, we applied all methods to identify enhancers in human heart left ventricle. Since data are available for only some of the epigenetic marks in this tissue, we retrained all methods to generate the enhancer predictions (see Supplemental Methods for more details). Then, we tested the enrichment of non-coding GWAS SNPs in these putative enhancers. Consistent with previous findings, only SNPs associated with traits in Cardiovascular category showed significant enrichment, indicating that the predicted enhancers are of reasonable quality (Figure 3.11A). However, we found that these SNPs were most enriched in REPTILE predicted enhancers, suggesting its better resolution and accuracy compared with other methods (Figure 3.11A-B).

Enhancers are expected to increase the transcription of target genes. To test this, we linked REPTILE putative enhancers to their target genes using expression quantitative trait loci (eQTL) data of left ventricle tissue from Genotype-tissue Expression (GTEx) Project (Supplemental Methods). We found that indeed genes linked to REPTILE enhancers showed significantly higher expression than genes linked to other genomic loci (Figure 3.11C).

3.4.7 REPTILE score correlates better with in vivo enhancer activity than open chromatin.

Although open chromatin signatures using DNase-seq[39]/ATAC-seq[40] were used for validation in this study, we found that REPTILE score is more predictive of the in vivo activity of DNA elements from VISTA database than open chromatin data (Figure 3.5A; Supplemental Methods). Two recent studies showed that low CG methylation in candidates of regulatory regions is an indicator of enhancers[41, 42]. To test this idea, we implemented an approach to predict enhancers based on the CG methylation level in DHSs (DHS+mCG; Supplemental Methods). Although useful, this approach does not provide better performance than REPTILE predictions (Figure 3.5A). We further tested other single histone marks as well as the H3K27ac signal in DHSs and found that none of these is as predictive as the REPTILE score (Figure 3.5A). Consistently, the enhancer predictions based on REPTILE score consistently achieved the best precision given different score cutoffs (Figure 3.5B-E; Supplemental Methods). These results highlight the value of a method that utilizes integrative data. At the same time, it suggests that open chromatin regions may not be the ideal data-type to validate predicted enhancers.

3.5 Discussion

In this study, we describe the development of a new algorithm, REPTILE, which is able to predict active enhancers by integrating tissue-specific histone modification data and base-resolution CG methylation (mCG) data. We found that the overall accuracy and resolution of REPTILE predictions exceeds other methods, especially when applied to cell/tissue types different than the training data. Further benchmarking revealed that REPTILEs performance is robust to different DMR inputs and reference choice (Figure 3.12 and Figure 3.13; Supplemental Note 1). In summary, by incorporating DNA methylation data produced by whole genome bisulfite sequencing and using information of cell/tissue type specific variation of epigenetic marks, REPTILE greatly improves upon current methods for annotation of enhancers across a variety of cell and tissue types (See also Figure 3.12 and Figure 3.14; Supplemental Note 2).

Although some methods showed better performance in a few tests, REPTILEs performance was superior in most tests. While we tried to evaluate the prediction accuracy of all methods in an unbiased manner, we should point out that these benchmarks might be further improved in several ways. First, the validated regions in mESCs were originally selected based on RFECS predictions, which introduces a potential bias. However, if this bias alters the performance of prediction algorithms it is likely to inflate the performance of RFECS more than

REPTILE. Second, the number of validated enhancer elements is currently limited, although this issue may be resolved in the near future, as more elements will be tested for in vivo function. Third, the negative data sets obtained from the VISTA enhancer database were mostly putative enhancer elements from previous studies and therefore may be very similar to true enhancers in many aspects, such as the degree of evolutionary conservation[43]. As a result, the prediction accuracy on VISTA enhancer dataset is likely to be lower than the accuracy of whole genome prediction because many of the negatives in the VISTA database actually have some enhancer-like characteristics, which likely makes them harder to differentiate from true positives. While improvements are possible (such as benchmarking of methods on genomic regions tested in high-throughput enhance assay and incorporating more sophisticated features in the REPTILE model), our results show that REPTILE outperforms existing enhancer prediction methods, especially for samples where training data is unavailable.

As epigenomic information of a larger number of cell/tissue types continues to be comprehensively profiled by the effort of Encyclopedia of DNA Elements (ENCODE)[32, 44, 45], Roadmap Epigenomics Mapping Consortium (REMC)[46], International Human Epigenome Consortium (IHEC) and other consortia, we envision that REPTILE will be a valuable tool to generate accurate enhancer annotations for these datasets, facilitating better regulatory DNA predictions and fueling new biological insights.

3.6 Methods

3.6.1 Overview of Data Acquisition.

In order to systematically benchmark REPTILE, we collected epigenomic data of various human and mouse cells and tissues. These epigenetic marks included base resolution DNA methylation data (WGBS) and six histone modifications: H3K4me1, H3K4me2, H3K4me3, H3K27ac, H3K27me3 and H3K9ac (Figure 3.6B-C; Supplemental Table S2 and S3). We downloaded data of five human cell lines from Xie et al.[47]: H1 human embryonic stem cells (H1), mesendoderm (MES), mesenchymal stem cells (MSC), neural progenitor cells (NPC) and trophoblast-like cells (TRO). Human data also contains WGBS of heart left ventricle from Schultz et al.[19] and histone modification data of the same tissue from Leung et al.[48]. In addition, we included data 9 mouse samples: mouse embryonic stem cells (mESCs) and 8 mouse tissues from E11.5 embryo (Supplemental Methods).

Next, to train the computational enhancer prediction methods, we obtained EP300 binding data from mouse and human ESCs (Supplemental Methods). It has been shown that EP300 binding is a key feature of a fraction of active enhancers but computational approaches are able to learn the chromatin signatures of these enhancers and predicts other active enhancers without EP300 binding[10, 11]. In this regard, we used EP300 binding sites as putative active enhancers in training datasets.

To validate the enhancer predictions from these methods, we collected in vivo enhancer validation data in E11.5 embryonic mouse tissues from the VISTA enhancer browser[31] as well as high-throughput report assay data in mESCs from Yue et al[32]. We also included in vivo validated embryonic heart enhancers from Narlikar et al[49]. In total, eight test datasets were used (Supplemental Figure 3.6D). In addition, in all five human cell lines, mESCs and 5 E11.5 mouse tissues, we downloaded publically available DNase-seq data to validate enhancer predictions, assuming the actual location of enhancers to coincide with distal DNase hypersensitivity sites (DHSs) in the corresponding cell/tissue types. See also Supplemental Methods for more details.

3.6.2 REPTILE

REPTILE (Regulatory Element Prediction based on TIssue-specific Local Epigenetic marks) is a novel algorithm, which generates high-resolution prediction of active enhancers genome-wide by integrating mCG and histone modification data. REPTILE uses the differentially methylated regions (DMRs) that are identified across all samples as high-resolution enhancer candidates and it is able to capture local epigenomic signatures that may otherwise be washed out in the signal of larger region. In addition, it takes into account the tissue-specificity of enhancers as features to further improve its performance; REPTILE predicts enhancers based on epigenomic data of not only the target sample (where enhancer predictions are generated) but also additional reference samples to exploit the useful information in variation between cells and tissues.

The overview of REPTILE workflow is shown in Figure 3.1C, which includes four major steps: 1) DMR calling: DMRs are identified by comparing the DNA methylomes of input samples. We first called differentially methylated sites

(DMSs). Next, we merged DMSs into blocks if they both show similar sample-specific methylation patterns and are within 250bp. These two steps were performed as previously described[19] (See also Supplemental Methods for details). We then filtered out the blocks that contain only one DMS. The remaining blocks were then extended 150bp from each side to include the two regions covered by first upstream and first downstream nucleosomes respectively. These extended blocks are defined as DMRs, which were used in later steps.

2) Data integration: Then, REPTILE integrates various types of input data to obtain the epigenomic signatures of DMRs and query regions, in preparing for the next two steps: enhancer model training and prediction generation. Specifically, each DMR or query region is represented as a feature vector and each variable in the vector corresponds to the intensity or intensity deviation of one epigenetic mark (Figure 3.1D). In this study, the intensity of each histone modification is defined as the log2 fold change RPM relative to control and the intensity of mCG is the CG methylation level. Note that different definitions of intensity can also be used, such as the RPM with subtraction of control or simply RPM of ChIP-seq itself. It makes REPTILE more flexible and allows various way of normalization to be imposed on the input data. Intensity deviation of an epigenetic mark is defined as the intensity in target sample subtracted by its mean intensity in reference samples (i.e. reference epigenome) and this type of feature quantifies the tissue-specificity of the epigenetic mark (Figure 3.6A). Since the data of reference samples is only used to calculate the mean signal value, REPTILE does not require that all epigenetic marks are available in all reference samples, i.e. missing data is allowed. However, the target sample, where enhancer predictions are generated,

must contain the data of all the epigenetic marks. In this study, we used seven epigenetic marks (DNA methylation and 6 histone modifications) and thus the complete REPTILE model contains in total 14 features (two features, intensity and intensity deviation, for each mark; Figure 3.14).

The input data varies according to the next step. 1) The training step requires data of known/putative enhancers (such as EP300 binding sites) and known negative regions as well as the DMR list and the epigenomic data of target sample and reference samples. 2) Prediction generation takes the enhancer model obtained from the training step, together with the DMRs the epigenomic data, as input. It also requires query regions. The query regions can be 2kb sliding windows with step size 100bp across the genome for generating genome-wide enhancer predictions (see below). They can also be pre-defined regions, such as conserved elements in the genome, where their enhancer activity is of interest. More details about REPTILE input preparation are available at https://github.com/yupenghe/REPTILE/.

3) Model training: In the next step, REPTILE enhancer model are trained by learning the epigenomic signatures of query regions, including known enhancers and negatives, as well as the DMRs within them. Specifically, one random forest classifier is trained to learn the epigenomic profiles of the labeled query regions, while another random forest classifier is trained to learn epigenomic features in the DMRs that overlap with the query regions. Both classifiers use same 14 features but the values of these features are calculated differently. The classifier for query regions computes feature values based on the epigenomic data of whole query regions, whereas the classifier for DMRs is trained and applied on the data of DMRs.

The random forest classifier for query regions can be trained on data of known active enhancers and negative regions. However, the classifier for DMRs cannot be trained in such straightforward way due to the lack of labels for DMRs. To circumvent this, we label all DMRs that are within known enhancers as active and we label the ones that are within negative regions as inactive. Then, we use these labels to train the random forest classifier for DMRs in a similar fashion as in the training of classifier for query regions. The rationale behind this is that (we assume that) DMRs within negative regions are inactive and part of the DMRs within active enhancers can be inactive. In the training dataset where negative regions greatly outnumber active enhancers, we expect that there are much more DMRs labeled as inactive than active. Therefore, although the inactive DMRs within active enhancers might be incorrectly labeled as active, they only compose a small portion of DMRs. In this paper, the ratio of negatives to positives in the training datasets is at least 7:1 (Supplemental Methods). The random forest model can be successfully trained on such data with a small fraction of instances incorrect labeled, which has been demonstrated by the better performance of REPTILE than existing methods. The implementation of random forest model is built on the R (version 3.2.1) package randomForest (version 4.6.12) with parameter ntree=2000, nodesize=1.

4) Prediction generation: Lastly, we apply the enhancer model learned in the training step to generate enhancer predictions. Specifically, for every query region or DMR, the corresponding random forest classifier will generate an enhancer confidence score, which is defined as the fraction of decision trees in the random forest model that vote in favor of the active enhancer class.

Given a set of regions of interest, REPTILE is able to predict their enhancer activity. First, REPTILE generates one enhancer confidence score based on the epigenomic signature of certain query region and also multiple scores based on the data of DMRs within it. Then, the maximum is assigned as the final score for this region. In this design, data of DMRs are used to complement the prediction based on query regions. We found that, with correct enhancer model, even if the DMRs were not correctly identified, the prediction performance did not decrease much (See REPTILE w/ shuf DMR in Figure 3.12). It is because the incorrect DMRs are not likely to show enhancer-like epigenomic signatures and low enhancer confidence scores will be assigned to them. In this case, the prediction will be dominated by the enhancer confidence score calculated based on the data of whole query regions (See REPTILE w/o DMR in Figure 3.12).

REPTILE can also generate enhancer predictions across the genome. In this study, we used REPTILE to first calculate enhancer scores for all DMRs in the genome as well as all 2kb sliding windows with 100bp step size across the whole genome. The empirical choices of window size 2kb and step size 100bp are based on the benchmark results from previous study[35, 50]. Then, DMRs with score higher than a given cutoff (0.5 is used in this study) are predicted to be enhancers (termed enhancer-like DMRs). In order to generate non-overlapping enhancer predictions, overlapping enhancer-like DMRs are merged into single prediction and it score is the highest score of all enhancer-like DMRs that are merged to form this prediction. Next, to capture the enhancers with no detectable mCG variation, REPTILE calls peaks of the enhancer scores across the sliding windows that pass the given score cutoff using the below procedure: 1.All sliding windows that pass the cutoff are labeled as enhancer candidates. Candidates that are within 1kb to each other are grouped into clusters. 2.For each cluster, the candidate with maximum score is set as a peak. If multiple candidates share the highest score, we randomly select one of them as the peak. 3.For each cluster, the peak and all candidates that are within 1kb of the peak are excluded from the candidate list. 4.Step 2 and 3 are repeated until the candidate list in each cluster is empty. After this process, all sliding windows that have score greater than threshold are either peaks or within 1kb to peaks. The rationale behind this is that the sliding windows adjacent to a peak are part of the peak. Lastly, the final predictions are the union of the enhancer-like DMRs and the sliding windows that are called as peaks but have no overlap with any enhancer-like DMRs. Similar to the prediction on given regions, this procedure is robust to incorrect DMRs because the enhancers that can be identified using the epigenomic mark of sliding windows will still be called.

3.6.3 Software availability

The REPTILE software is published under the BSD 2-Clause License. It was written in R and python. The R code was submitted as an independent R package, called REPTILE, in the Comprehensive R Archive Network (CRAN). The source code, supplemental dataset, pre-trained enhancer models, usage and further details of the complete pipeline are available in https://github.com/yupenghe/REPTILE.

3.7 Acknowledgements

We thank Dr. John A. Stamatoyannopoulos for generously sharing the DNase-seq data of five E11.5 mouse tissues. We specifically thank Dr. Nisha Rajagopal for kindly helping with the data from MERA. We thank Drs. Shaoshan Carol Huang, Chongyuan Luo, and Manoj Hariharan for their critical comments. Y.H. is supported by the H.A. and Mary K. Chapman Charitable Trust. D.U.G is supported by the A.P. Giannini Foundation and NIH IRACDA K12 GM068524. Transgenic mouse work was conducted at the E.O. Lawrence Berkeley National Laboratory and performed under Department of Energy Contract DE-AC02-05CH11231, University of California. J.R.E is an Investigator of the Howard Hughes Medical Institute and is supported by grants from the Gordon and Betty Moore Foundation (GBMF3034), the National Institutes of Health (R01 MH094670 and U01 MH105985), and the California Institute for Regenerative Medicine (GC1R-06673-B). This work is funded by National Institutes of Health (U54 HG006997).

Chapter 3, in full, is a reprint of the material as it appears in Proceedings of the National Academy of Sciences 2017. Yupeng He, David U. Gorkin, Diane E. Dickel, Joseph R. Nery, Rosa G. Castanon, Ah Young Lee, Yin Shen, Axel Visel, Len A. Pennacchio, Bing Ren, and Joseph R. Ecker. Improved regulatory element prediction based on tissue-specific local epigenomic signatures. Proceedings of the National Academy of Sciences. 28;114(9):E1633-1640, 2017 Feb. http://www.pnas.org/content/114/9/E1633 The dissertation author was a primary investigator and author of this paper.

3.8 Contributions

Y.H. conceived and designed the algorithm, performed all computational experiments and wrote the manuscript. D.U.G, B.R. and J.R.E edited the manuscript. A.V.s group and L.A.P.s group collected the tissues from E11.5 mouse embryo, which were later profiled for epigenetic marks. D.D., A.V. and L.A.P conducted transgenic mouse assay and generated the newly validated VISTA enhancers. D.U.G, A.Y.L, Y.S. and B.R. generated the histone modification data for the E11.5 tissues. J.R.N and R.C generated the whole genome bisulfite sequencing data for the E11.5 tissues. J.R.E supervised the project.

3.9 Disclosure Declaration

The authors declare no conflict of interest.

3.10 Supplemental Notes

3.10.1 Performance of REPTILE is robust to choice of reference and suboptimal DMR calling.

Compared with other methods, REPTILE utilizes DMRs to improve prediction resolution. However, there is no consensus DMR definition and different algorithms may identify different regions as DMRs[51]. We test the robustness of REPTILE to DMR input using mouse data because experimentally validated enhancers are available in mouse samples (Figure 3.6D). First, we shuffled the genomic location of DMRs and used these shuffled DMRs in the prediction step, whereas the enhancer model was learned using unshuffled DMRs (REPTILE w/ shuf DMR) (see Supplemental Methods). We found that the prediction accuracy remained superior to existing methods in 4 out of the 8 test datasets (Figure 3.12A). As expected, without meaningful DMRs, this method has worse prediction resolution than REPTILE with complete input set (Figure 3.12B and D). We also found that fewer predictions were near distal DHS compared to REPTILE with full inputs (Figure 3.12C and E). However, the REPTILE w/ shuf DMR predictions remain comparable with existing methods, indicating that REPTILEs performance is robust to suboptimal DMR input given a correctly pre-trained enhancer model. Inspired by these results, we provide pre-trained enhancer models along with the software to facilitate the use of REPTILE. This robust performance is likely due because REPTILE generates good enhancer prediction solely based on epigenomic signatures of the query regions (REPTILE w/o DMR) and the DMR input simply improves upon an already relatively accurate prediction.

Next, we asked whether the performance of REPTILE is robust to the choice of reference samples. To test this, we ran REPTILE with a different strategy of choosing the reference (Figure 3.13). Instead of using all non-target samples as the reference, we only used mESCs, E11.5 Craniofacial and E11.5 Liver as reference in the prediction step REPTILE alt Ref). We then evaluated this strategy using mouse data. Specifically, we ran REPTILE to predict the enhancer activity of elements from VISTA enhancer browser in six E11.5 tissues (Figure 3.13A). For each target sample, we only used data of the target sample, mESCs, E11.5 Craniofacial and E11.5 Liver. We trained an enhancer model for each target sample on data of mESCs with the target sample, E11.5 Craniofacial and E11.5 Liver as reference (Supplemental Methods). In the prediction step, mESCs, E11.5 Craniofacial and E11.5 Liver were used as the reference (Supplemental Methods). For each target sample, DMRs were called across methylomes of the four samples. We found that even if we changed the reference, REPTILE (REPTILE alt Ref) showed performance as good as the previous setup (REPTILE; Figure 3.13). We further used DHS data to validate the enhancer predictions from these two setups and they showed similar prediction accuracy and resolution. Collectively, these results demonstrate that REPTILEs performance is robust to different reference choice.

3.10.2 Epigenomic variation information improves enhancer prediction resolution and accuracy.

Use of the random forest algorithm allowed us to identify key epigenetic features in the enhancer prediction model (Figure 3.14; Supplemental Methods). In the mouse enhancer model, we found that mCG was the most informative feature for predicting enhancer activities of DMRs, while H3K27ac was the most predictive mark for query regions. This is likely due to the fact that hypomethylation tends to be restricted within DMRs, and thus becomes less predictive in larger query regions where hypomethylation pattern is washed out. In the human enhancer model trained on H1 cells, H3K4me2 is the most informative feature in both classifiers. We also found several other high-ranking features including intensity deviation features, such as H3K4me2-dev and H3K27me3-dev, indicating the necessity to capture the tissue-specificity of epigenetic marks for enhancer prediction (Figure 3.14). When the intensity deviation features were removed (REPTILE w/o

Ref), REPTILE prediction accuracy decreased, even though the results remained comparable or superior to other methods (Figure 3.12).

Next, to understand the contribution of DMRs, we tested REPTILE without DMR input (REPTILE w/o DMR). We found that the midpoint genomic locations of these predictions were 30-40bp further from the closest distal DHS compared to the predictions made by REPTILE with DMR input (Figure 3.12B and D). Also, the percentage of DHS-supported predictions slightly decreases (Figure 3.12C and E). However, in the enhancer validation datasets, the prediction accuracy without DMR input remains as good as the REPTILE method will all inputs (Figure 3.12A). These results indicate that the inclusion of tissue-specificity information improves prediction accuracy while DMRs are necessary for more refined prediction of enhancer locations.

3.11 Supplemental Methods

3.11.1 Whole-genome Bisulfite Sequencing Data

The raw reads of MethylC-seq or whole-genome bisulfite sequencing (WGBS) data of eight mouse tissues from E11.5 embryo were downloaded from the ENCODE website (https://www.encodeproject.org/). Mouse embryonic stem cells (mESCs) WGBS data was obtained from Gene Expression Omnibus (GEO). The accession numbers of the two mESCs replicates are GSM1162043 and GSM1162044. The paired-end data (GSM1162045) of the second replicate was not included to avoid potential bias due to different data type (paired-end versus single-end). WGBS raw reads of human cell lines, H1 human embryonic stem cells (H1), mesendoderm (MES), mesenchymal stem cells (MSC), neural progenitor cells (NPC) and trophoblast-like cells (TRO), were obtained from SRA (accession SRP000941). For MSC, whose methylome had been sequenced in paired-end, we mapped the first read in each pair to avoid problems in processing overlapping reads similar to Schultz et al.[19]. WGBS data of human heart left ventricle was downloaded from GEO (GSM983650). The sources of all the WGBS data can be found in Supplemental Table S2.

WGBS data were processed as previously described[52], using mm10 reference for mouse data and hg19 reference for human data. Only autosomes, sex chromosomes, mitochondrial chromosomes and the genome sequence of lambda phage (as control) are included in the reference genome. The sequences were downloaded from UCSC genome browser[53]. For each sample, if biological replicates were available, the data of replicates were combined. To quantify the methylation landscape, we divided the genome into 100bp bins and calculated the (weighted) methylation level[54] for each bin. Weighted methylation level is also called as the CG methylation (mCG) intensity and it is defined as the ratio of the sum of methylated basecall counts over the sum of both methylated and unmethylated basecall counts across all CG sites in a given region[54]. For each sample, these values were used to generate a file (in bigWig format) to store the methylation levels of all bins (see also https://genome.ucsc.edu/goldenpath/help/bigWig.html for more about bigWig format). In all the analyses in this paper, the methylation levels of any region was obtained from these bigwig files using the bigWigAverageOverBed executable from UCSC genome browser[53].

3.11.2 Identification of Differentially Methylated Region (DMRs)

DMR calling was done using very similar procedure as Schultz et al.[19]. We included (and rephrased) its the entire description here and highlighted the modifications we made. In the procedure, we considered bisulfite sequencing as a binomial process and defined a stochastic model in which at each position, the observed number of reads supporting methylated cytosine in each sample is drawn from a binomial distribution. The true fraction of methylated alleles in the population in given sample at given cytosine in CG context, x_n^i, is the parameter of the binomial distribution, where i denotes the position of cytosine and n denotes the sample. The null hypothesis is that the methylation level (x_n^i) at this position is equal across all samples: $x_n^i = x^i$ for all n.

Our procedure is designed to test whether the observed data are consistent with the null hypothesis, or alternatively if there is a significant deviation from equal methylation levels. To do this, we compute a goodness-of-fit statistic, s, introduced by Perkins et al[55]. We arrange the observed data in an $Nx2$ table, with each row for each of the N samples and the two columns for the number of reads supporting methylated and unmethylated cytosines respectively. The number of observed reads in sample n at position i is o_{nj}^i, where $j = 1$ for methylated reads and $j = 2$ for unmethylated reads. The expected number of reads in sample n with methylation state j under the null hypothesis is e_{nj}^i:

$$e_{nj}^i = (\sum_{m=1}^{N} o_{mj}^i)(\sum_{k=1}^{2} o_{nk}^i)/M^i \tag{3.1}$$

where $M^i = \sum_{n=1}^{N} \sum_{k=1}^{2} o^i_{nk}$ is the total number of reads in all samples. The statistic for the goodness of fit is

$$s^i = \sqrt{\frac{1}{2N} \sum_{n=1}^{N} \sum_{j=1}^{2} (o^i_{nj} - e^i_{nj})^2} \tag{3.2}$$

Next, we simulated read count data under our stochastic model assuming the null hypothesis in the following way:

- Set all cell counts in the table to zero
- Randomly select a cell in the table with probability equal to the expected counts divided by the total number of counts in the table $(\frac{e^i_{nj}}{M^i})$. Increment the value in this cell by one.
- Repeat this procedure M^i times.
- Finally, calculate the value of the statistic, s^i_{shuff}, for the randomly generated table.

This randomization procedure was repeated until we observed 100 iterations with a value of s^i_{shuff} that was at least as extreme as that of the observed data, s, up to a maximum of 3,000 iterations. The p-value at position i was then computed as:

$$p^i = \frac{R^i + 1}{T^i} \tag{3.3}$$

Where R^i is the number of randomization where a statistic greater than or equal to the original tables statistic was observed. T^i is the total number of randomizations that were conducted. Our adaptive permutation procedure ensures that any sites which we may potentially identify as significantly differentially methylated with $p^i < 0.01$ will be sampled 3,000 times. At other sites, we have observed an appreciable number (100) of permutations more extreme than our original test statistic ($s \geq s_{shuff}$) and the p-value for these sites will be $p \geq (100 + 1)/3000 = 0.034$; these sites will therefore not be called as differentially methylated.

To control the false discovery rate (FDR) at our desired rate of 1%, we used a procedure designed for permutation-derived p-values[56]. First we generated a histogram of the p-values across all cytosines in CG context as described before. Next, we calculated the expected number of p-values to fall in a particular bin under the null hypothesis. This expected count is computed by multiplying the width of the bin by the current estimate for the number of true null hypotheses (m_0), which is initialized to the number of tests performed. We then identified the first bin (starting from the most significant bin) where the expected number of p-values is greater than or equal to the observed value. The differences between the expected and observed counts in all the bins up to this point are summed, and a new estimate of m_0 is generated by subtracting this sum from the current total number of tests. This procedure was iterated until convergence, which we defined as a change in the m_0 estimate less than or equal to 0.01. With this m_0 estimate, we were able to estimate the FDR corresponding to a given p-value cutoff by multiplying the p-value by the m_0 estimate (the expected number of positives at that cutoff under the null hypothesis) and dividing that product by the total number of significant tests we detected at that p-value cutoff. We chose the largest p-value cutoff that still satisfied our FDR requirement.

Next, we combined significant sites (differentially methylated sites or DMSs) into blocks if they were within 250bp and showed methylation changes in the same direction (e.g. sample A was hypermethylated and sample B was hypomethylated at both sites). A sample was considered hypo- or hyper-methylated if the deviation of observed counts from the expected counts was in the top or bottom 1% of deviations. These residuals were calculated for a position i using the following formula for a given cell in row n and column j of the table:

$$\frac{o^i_{nj} - e^i_{nj}}{\sqrt{e^i_{nj} * (1 - \sum_{m=1}^{N} \frac{e^i_{mj}}{M^i}) * (1 - \sum_{k=1}^{2} \frac{e^i_{nk}}{M^i})}} \tag{3.4}$$

The distinction between hypermethylation and hypomethylation was made based on the sign of the residuals. For example, if the residual for the methylated read count of sample A was positive, it was counted as hypermethylation. Furthermore, blocks that contained fewer than 2 DMSs were discarded. Instead of the 10 DMS cutoff used in original procedure, we used a more lenient 2 DMS cutoff to get a more comprehensive list of DMRs (enhancer candidates) to feed REPTILE. As an additional step to the original procedure, we next extended the remaining blocks by 150bp from both side and defined them as DMRs. The purpose of this extra step is to include regions where the histone modifications generally occur the upstream and downstream nucleosomes flanking putative enhancers (typically nucleosome-free regions).

3.11.3 Applying the DMR calling algorithm on human and mouse cells and tissues

To obtain DMRs for mouse samples, we applied the above calling algorithm on the mCG profiles of mESCs and eight E11.5 mouse tissues. In total, 542,139 DMRs were identified, with average length 484bp and covering over 262Mb or 10% of the genome. We found that 97% of the experimentally validated enhancers (246 out of 253) in VISTA enhancer browser[31] overlap with DMRs. By contrast, out of the 45 elements in VISTA enhancer browser that did not overlap with any DMRs, 38 (86%) did not show any enhancer activity, implying that differential methylation is a significant enhancer signature.

We applied the same procedure to call DMRs across the mCG profiles of all human cell lines. We identified 159,474 DMRs and their average length is 439bp. These DMRs covered 2% of the genome.

3.11.4 Chromatin and Transcription Factor ChIP-seq Data

For the eight E11.5 mouse tissues, we downloaded the ChIP-seq data of six previously identified enhancer-related histone marks (H3K4me1, H3K4me2, H3K4me3, H3K27ac, H3K27me3 and H3K9ac) and the corresponding control from the ENCODE project website (https://www.encodeproject.org/). For mESCs, ChIP-seq data of the same histone modifications and the corresponding controls were downloaded from GEO (Supplemental Table S3). In addition, ChIP-seq data of EP300 and its corresponding control data were downloaded from GEO (Supplemental Table S3). We also downloaded ChIP-seq data of 12 transcription factors in mESCs from GEO (Supplemental Table S3).

All mouse ChIP-seq data were processed using the ENCODE uniform processing pipeline for ChIP-seq: First, reads were mapped to the mm10 reference using bwa[57] (version 0.7.10) with parameters -q 5 -l 32 -k 2. The mm10 reference only contains autosomes, sex chromosomes and mitochondrial sequences. It is also called as mm10-minimal in the ENCODE website. Then, Picard tool (http://broadinstitute.github.io/picard/, version 1.92) was used to remove PCR duplicates using parameter REMOVE_DUPLICATES=true.

For chromatin ChIP-seq data of human cell lines and heart left ventricle tissue, we directly downloaded the alignment files (labeled as Unconsolidated Epigenomes (Uniform mappability)) from the data portal of the NIH Roadmap Epigenomics Mapping Consortium (http://egg2.wustl.edu/roadmap/web_portal/index.html). We obtained ChIP-seq data of H3K4me1, H3K4me2, H3K4me3, H3K27ac, H3K27me3, H3K9ac and corresponding control for all five human cell lines. For heart left ventricle, we downloaded H3K4me1, H3K4me3, H3K27ac, H3K27me3 and control since other histone marks were not available.

For each histone modification mark in human and mouse samples, we represented it as continuous enrichment values of 100bp bins across the genome. Specifically, we first extended reads to 300bp (expected fragment length) using the -r option (along with -s and -l 0 options) in slopBed from bedtools[58]. We then divided the mouse genome into 100bp bins and for each bin, we calculated log2 fold RPM relative to control. RPM for control experiment in each bin is smoothed by averaging it over the RPMs of 2 bins upstream and 2 bins downstream. RPM (Reads Per Million mapped reads) for a given bin is defined as the number of mapped reads that overlap (1bp) with the bin divided by the total number (in million reads) of the uniquely mapped reads in the genome.

For the ChIP-seq data of TFs and EP300 in mESCs, we used MACS[59] (1.4.2) to call peaks with default parameters. The reported TF peaks were filtered out if they are within 1kb to any transcription start sites (TSSs) of genes in mouse GENCODE[60] annotation (M2).

3.11.5 EP300 and Transcription Factor Binding Sites in H1

We downloaded the binding sites of DNA-binding proteins in H1 from ENCODE data portal in the UCSC genome browser (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegTfbsClustered/wgEncodeRegTfbsClusteredWithCellsV3.bed.gz). The binding sites of EP300 were used as positive instances (i.e. putative active enhancers) in the training of enhancer prediction methods. The distal binding sites of remaining DNA-binding proteins, excluding CTCF, were used to validate the prediction in H1 in Figure 3.2A (See later section for details). The reason to exclude CTCF was that CTCF played a major role in shaping the chromatin architecture and its binding sites included insulators[61]. Distal binding sites are at least 1kb away from any TSSs in the human GENCODE annotation (release 19).

3.11.6 Enhancer Validation Data

In order to evaluate the enhancer prediction accuracy, we collected publicly available data of experimentally validated enhancers and negative sequences (sequences that showed no detectable enhancer activity) from three sources (Figure 3.6D). The *in vivo* and *in vitro* data were used to construct the 8 test datasets used in benchmark (Figure 3.6D).

- From Yue et al.[32], we downloaded 212 regions that were tested for in vitro enhancer activity by luciferase reporter assay in mESCs. The original coordinates of these regions were in mm9 reference and they were liftover to mm10 using liftOver utility from UCSC genome browser[53]. One region was filtered out in this process. Out of the remaining 211 tested regions, 131 showed enhancer activity in mESCs and were labeled as positive, while the rest were labeled as negative.

- In addition, we obtained in vivo enhancer validation data from VISTA enhancer browser[31] (Oct 24th, 2015). In total 546 mouse sequences were tested for in vivo enhancer activity in E11.5 mouse embryo using transgenic reporter assay. Their mm9 coordinates were liftover to mm10 and one region was removed. In the eight E11.5 mouse tissues where epigenomic data is available, six of them had reasonable number (¿=30) of validated enhancers. We used the data of these tissues (forebrain, midbrain, hindbrain, heart, limb and neural tube) to build six test datasets (Figure 3.6D). In this study, we only included the mouse sequences in VISTA database and excluded all human sequences. The rationale is that the in vivo enhancer activity of human sequences may be different from the activity of their mouse counterparts (orthologs), preventing them from being good validations.

- We also included 36 in vivo validated sequences that were tested in vivo in the heart of zebrafish embryo from Narlikar et al[49]. The enhancer activity in the embryonic heart of zebrafish was shown to be conserved in mouse embryo[49]. Based on this, we used these regions as approximation of enhancers in E11.5 mouse heart. The original dataset included 46 regions with coordinates in hg18 human reference genome. The hg18 coordinates were first liftover to hg19, which were then converted to mm10. In this process, 10 regions were eliminated and the remaining 36 were included in later analysis.

3.11.7 DNase-seq Data

The DNase Hypersensitivity Sites (DHSs) identified based on DNase-seq data were used to validate enhancer predictions. DHS calls of all five human cell lines were obtained from the NIH Roadmap Epigenomics Mapping Consortium (http://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidated/narrowPeak/). We downloaded the narrow DHS peaks from MACS2[59] (files whose names ended with -DNase.macs2.narrowPeak.gz).

DHS calls of mESCs were downloaded from UCSC genome browser (http://hgdownload.cse.ucsc.edu/goldenPath/mm9/encodeDCC/wgEncodeUwDgf/wgEncodeUwDgfEscj7129s1ME0PkRep1.narrowPeak.gz). The coordinates of these elements (mm9) were liftover onto mm10. DNase-seq data and DHSs in E11.5 mouse tissues was downloaded from the ENCODE project website (https://www.encodeproject.org/). We found that the DNase-seq data were available for five

E11.5 tissues. The tissues and the corresponding accessions in the ENCODE project website are E11.5 craniofacial (ENCSR196VDE), E11.5 neural tube (ENCSR312QVY), E11.5 midbrain (ENCSR292QBA), E11.5 hindbrain (ENCSR358ESL) and E11.5 limb (ENCSR661HDP). Narrow peak files were downloaded and each peak call was defined as one DHS.

DNase-seq data was available for two biological replicates of each mouse E11.5 tissue. The DHSs of two biological replicates were combined using bedops (http://bedops.readthedocs.io/en/latest/content/usage-examples/ master-list.html)[62]. Below is the procedure description adapted from the text in the bedops webpage. The procedure starts with the union of DHSs called in both replicates (i.e. original elements) and an empty master list, which stores the final result.

1. Original elements not yet in the master list are merged into non-overlapping intervals (using bedops -m).
2. For each merged interval, the original element of highest score within the interval is selected to go into the master list.
3. Any original elements that overlap the selected element are thrown out.
4. Repeat the step 1, 2 and 3 until no original element is left. Then the master list is reported as the final DHS list.

3.11.8 Existing Enhancer Prediction Approaches

To evaluate the performance of REPTILE, it was compared to four publicly available methods, PEDLA[34], RFECS[35], DELTA[36] and CSIANN[37]. All of these methods are supervised approaches, meaning that they learned the profiles of enhancers from data with labels and then classify regions with no labels. Specifically, they first represent genomic regions using histone modification data (and maybe other data types). Then, machine learning technique is used to learn the histone modification signatures of (putative) enhancers and background regions. Last, the trained computational model is used to classify unknown regions into enhancers or negative regions.

Their differences lie in the distinct strategy used to represent genomic regions and their different underlying machine learning framework.

- PEDLA used the histone modification signals and evolutionary conservation score as low-level features and it is capable of incorporating additional data types. Then, PEDLA applies Deep Neural Network (DNN), in an unsupervised fashion, to extract high-level features from these low-level features in all 200bp non-overlapping bins across the genome. Lastly, the DNN is used to learn the feature signatures of enhancers and background sequencers (supervisedly) and then makes predictions.

- RFECS represents the shape and intensity of each histone modification (ChIP-seq) signal in each 2kb genomic window using a feature vector of length 20. Specifically, RFECS divides the 2kb window equally into 20 100bp non-overlapping bins and the values in the feature vector correspond to the signal values in the 20 100bp bins. Next, a random forest classifier[33] with some modification on the node separator is trained on this type of data on putative enhancers and background sequences. This model is then used to delineate enhancer-like chromatin signatures from genomic background.

- DELTA defines four shape features to describe the histone modification (ChIP-seq) signature and then uses AdaBoost algorithm[63] to distinguish enhancers from negative regions based on this representation schema.
- CSIANN was built on neural network framework and it makes predictions based on the histone modification signals of 2kb non-overlapping genome windows.

3.11.9 Running REPTILE and Existing Enhancer Prediction Methods

REPTILE and the four existing methods were trained in mESCs (for mouse enhancer prediction) or in H1 (for human enhancer prediction) by learning the epigenomic signatures of known/putative enhancers (EP300 binding sites) and negative regions (promoters and genomic background). The promoters are defined as 2kb regions around TSSs and the TSSs were based on GENCODE annotation (mouse - M2; human - release 19). The enhancer predictions are provided as part of Supplemental Data.

- **REPTILE**: The training dataset for REPTILE was constructed using a similar strategy used for training RFECS previously[35]. The training dataset for mouse enhancer prediction is composed of 5,000 positive instances (enhancers) and 35,000 negatives (negative regions). Positives were the +/- 1kb regions around the summits of top 5,000 EP300 peaks in mESCs. Negatives included 5,000 randomly selected promoter regions and 30,000 (6 times than number of positives) randomly chosen 2kb bins. The 2kb bins have no overlap with promoters, top 5,000 EP300 binding regions or any regions in the mESCs test dataset. The training dataset for human enhancer prediction was constructed similarly. It includes 5,476 distal EP300 binding sites in H1 as positives and equal number of randomly chosen promoters and 32,856 (6 times than number of positives) 2kb bins. Score cutoff 0.5 was used to generate genome-wide enhancer predictions for both human and mouse samples.

- **PEDLA**: The training dataset for PEDLA were constructed similarly as REPTILE. The only difference is that the number of 2kb bins is 9 times of the number of positives to be consistent with how PEDLA was trained[34]. We benchmarked various parameters of PEDLA and found that single layer with 500 neurons performed well in both human and mouse data (data not shown). This setting was used for running PEDLA. In the current implementation of PEDLA, hidden markov model (HMM) is used to generate the final enhancer prediction based on the scores from the artificial neural network model. Score is defined as the observatory probability conditioned on enhancer state divided by prior probability of enhancer state (i.e. base rate). However, its performance was not as good as other methods (labeled as PEDLA (HMM) in Figure 3.7A-B). Therefore, we implemented an alternative enhancer calling approach by applying the peak-calling algorithm used by REPTILE on the scores model. We called this approach PEDLA in this study. It showed better performance than the current PEDLA implementation (Figure 3.7A-B). Score cutoff 5 was used to generate enhancer calls.

- **RFECS**: RFECS were trained on the same dataset as REPTILE. The default cutoff 0.5 was used to generate genome-wide enhancer predictions in mESCs and all human cell types. In E11.5 tissues, we used cutoff 0.2 to ensure that the number of putative enhancers was practically useful and enough (¿10,000) for validation.

- **DELTA**: For mouse enhancer prediction, the training dataset for DELTA were composed of the top 5,000 EP300 binding sites in mESCs and all promoters in mouse genome. For human enhancer prediction, the training dataset includes the 5,476 EP300 binding sites in H1 and all promoters in human genome. In the step of generating genome-wide predictions, we switched to the peak-calling algorithm used by REPTILE. It is because the default peak-calling algorithm in DELTA does not consider the spacing between peaks and thus generates a large number of predictions within 100bp to each other, which is not desirable in practice. Score cutoffs 0.1 in mESCs and human samples, whereas 0.05 in E11.5 tissues were used to generate enough genome-wide predictions for validation.

- **CSIANN**: For mouse enhancer prediction, top 500 EP300 binding sites in mESCs and gene annotation from GENCODE (M2) were used as input for CSIANN training. Similarly, for human enhancer prediction, top 500 EP300 binding sites in H1 and gene annotation (human GENCODE release 19) were used for training. The small number of positives in the input was due to the fact that current CSIANN implementation imposed a size limit on the training data. Default settings were used for both model training and prediction generation.

3.11.10 Evaluating the Performance of Methods using Cross Validation

In mESCs and H1, we used cross validation to evaluate the performance of each method similar to Liu et al.[34]. Results are shown in Figure 3.7A-B. The training data for PEDLA was used since it contains the most regions. We used 5-fold stratified cross validation, in which the ratio of positives to negatives was maintained in each round. Note that the current implementation of RFECS, DELTA and CSIANN did not allow users to specify the negative regions for training. Therefore, we just changed the positives for training RFECS and DELTA in cross validation, whereas we used the top 500 positives to training CSIANN due to its limit in current implementation. In addition to these methods, we also included the chromatin states of mESCs and H1 (if available). The chromatin state of H1 were downloaded from the ENCODE portal at UCSC genome browser. The chromatin state of mESCs was downloaded from github (https://github.com/gireeshkbogu/chromatin_states_chromHMM_mm9/blob/master/). Strong enhancers in the chromatin state map were regarded as enhancer predictions.

To ensure a fair comparison, we selected equal number of predictions from each method (if possible) and then resized them to 2kb regions while maintaining their center. Predictions from REPTILE, PEDLA, RFECS, DELTA and CSIANN were ranked and the top ones were selected. Since the enhancer predictions from ChromHMM[64] and Segway[65] cannot be ranked, we randomly chosen the same number of putative strong enhancers from their annotations.

To evaluate the prediction results, we first defined: True positives (TP) are positives that are overlapped with predicted enhancers. False positives (FP) are negatives that are overlapped with predicted enhancers. True negatives (TN) are negatives that do not overlap any enhancer predictions. The remaining are false negatives (FN), which are positives that are not predicted as enhancers. Next, we calculated the below metrics for the predictions from each methods:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision = TP / (TP + FP)

Recall/Sensitivity = TP / (TP + FN)

Specificity = TN / (TN + FP)

GM (geometric mean) = sqrt(Sensitivity * Specificity) where sqrt is square root.

F1-score = 2 / (1 / Precision + 1 / Recall)

DHS is the fraction of enhancer predictions that are overlapped with DHS but not any TSSs.

TFBS is the fraction of enhancer predictions that are overlapped with distal TFBSs but not any TSSs.

Misclassification is the fraction of enhancer predictions that are overlapped with any TSSs.

3.11.11 Evaluating the Prediction Accuracy on Data of Validated Enhancers

We also validated the predictions using experimentally validated regions; we applied all the methods to predict the enhancer activity of tested regions in the eight test datasets, which contain validated enhancers, and negative regions (Figure 3.6D): First, we ran all methods to generate scores for 2kb sliding windows in the genome. Then, the score of each tested region is assigned as the score of the sliding window whose center is the closest to the center of the tested region. If the centers of two sliding windows are equally close to the center of one tested region, the maximum score is used.

The reason behind this procedure is that RFECS, DELTA and CSIANN were designed to predict the enhancer activity of 2kb sliding windows in the genome and their current implementations were unable to calculate scores for pre-defined regions. The strategy of test PEDLA was different because it made predictions based on the chromatin profiles of 200bp bins, which is much smaller than 2kb. To address this issue, for each 2kb sliding window, we used the maximum PEDLA score among scores of overlapping 200bp bins as the score of the 2kb window. Also, to ensure the prediction results from all methods are comparable, we chose to run REPTILE to predict enhancer activity of 2kb sliding windows in the genome as well: REPTILE will first generate multiple enhancer confidence scores for each 2kb sliding window based on the epigenomic signature of the whole region as well as that of the DMRs within the region and then the highest is assigned as the final score for the window.

Then, the Area Under the Precision-Recall curve (AUPR) was used to measure the performance of each method in the test datasets. Precision is defined as the fraction of predictions that are real enhancers, i.e. (True positives) / (True positives + False positives). Recall is defined as the percentage of real enhancers that are predicted as positive, i.e. (True positives) / (True positives + False negatives). Precision-Recall curve can be drawn by changing the score cutoff. AUPR is defined as the (area) integral between the curve and two axes. R package flux (0.3.0) was used to implement the calculation of AUPR.

3.11.12 Validating Enhancer Prediction with distal TFBSs and distal DHSs

We overlapped the mESCs and H1 predictions with the distal DHSs and the distal transcription factors binding sites (TFBSs). We calculated the distance between the center of each prediction and the closest distal DHS (or the closest distal TFBSs). If the distance is no greater than 1kb, we see it as an overlap. Similar analysis was done to measure the overlaps with TSSs. If the center of certain prediction is within 1kb to any TSSs, it was counted as overlapping. Based on the overlap patterns, we divided the mESCs predictions into 5 categories: TSS proximal (overlap with TSSs), DHS (overlap with distal DHS only), TFBS (overlap with distal TFBS only), TFBS+DHS (overlap with both distal DHS and distal TFBS) and Unknown (none of the above). If a prediction is within 1kb to any TSSs, it will be consider as TSS proximal regardless of its distance to DHSs or TFBSs. TFBS, DHS and TFBS+DHS are considered as true positives, whereas TSS proximal is considered as false positive.

3.11.13 Validating Enhancer Prediction using MERA identified regulatory elements

To validate the enhancer predictions using various source of evidence, we also acquired the data of regulatory elements identified by a genome mutation screening approach, MERA (multiplexed editing regulatory assay)[38]. Briefly, GFP was knocked in to a selected gene and then CRISPR-Cas9 system was used to disrupt regions that are likely to have regulatory function on the selected gene. Next, the targeted regions of the guide RNA (gRNA) that significantly reduced the GFP signal were identified as regulatory elements. We downloaded the data from previous publication[38], where MERA assay was conducted in mESCs on four genes, *Tdgf1*, *Zfp4*, *Nanog* and *Rpp25*, separately. We used the same procedure as in the publication[38] to select gRNAs that were statistically significantly overexpressed in GFP-negative cells. Only the gRNAs that showed significance in all replicates were considered. Next, we merged the targeted regions of these gRNAs if they were within 100bp and we then filtered out the merged elements that were within 1kb to any TSSs. Then, we overlapped the (top 35k) mESCs enhancer predictions from each method with these distal merged elements. Last, we calculated the percentage of the distal merged elements that were within 500 bp to the center of any enhance predictions (Figure 3.7C). The final distal merged elements are available in Supplemental Data.

3.11.14 Evaluation of Genome-wide Enhancer Predictions

We evaluated the quality of genome-wide enhancer predictions by measuring the fraction of predictions that show evidence of distal open chromatin, how close the predictions are to nearest distal open chromatin regions (DHSs) and the percentage of predictions that are more likely to be (misclassified as) promoters. Before calculating these metrics, we selected the same number of predictions from each method to ensure a fair comparison. In human cell lines, the top 20,000 predictions were considered, which is similar to the strategy used in a recent study[34]. In mESCs, the top 35,000 putative enhancers were selected. In E11.5 tissues, the top 10,000 were selected because generally fewer predictions were generated in these samples than in mESCs. In total, three metrics were calculated.

- First, we measured the fraction of predictions whose centers were within 1kb to distal DHSs (1kb from any TSSs) and were at least 1kb away from any TSS. We called this metric as validation rate.

- In addition to the validation rate, we calculate a metric misclassification rate as the fraction of predictions that are within 1kb to TSS. These predictions are likely to be promoters and thus are misclassified.

- Furthermore, we measured the average distance between the centers of predictions and distal DHSs if the distance is no greater than 1kb. We intend to use this metric to measure the resolution of predictions and the ability of the method to accurately locate enhancer regions with little influence by false positives. Therefore, the predictions whose centers are 1kb away from DHS were not included in the calculation as they are considered as false positives.

3.11.15 Transgenic mouse experiments

Enhancer names (mm and hs numbers) are the unique names used in the VISTA Enhancer Browser (http://enhancer.lbl.gov/)[31]. Enhancer sequences were amplified from human (hs numbers) or mouse (mm numbers) genomic DNA and cloned into an hsp68-lacZ expression vector[66]. Genome coordinates and primer sequences for all elements are listed in Supplemental Table S1. Transgenic mouse assays were performed as previously described[66, 67] in Mus musculus FVB strain mice. All animal work was reviewed and approved by the Lawrence Berkeley National Laboratory Animal Welfare and Research Committee.

We then overlapped these newly validated VISTA enhancers with REPTILE predictions in E11.5 tissues. The murine VISTA elements (mm9) were lifted to mm10 using minMatch=0.95 using liftOver, while the human ones (hg19) were lifted to mm10 using minMatch=0.10. The resulting mm10 coordinates were intersected with the REPTILE predictions in E11.5 tissues and elements that were overlapped by at least 1bp were reported.

3.11.16 TF-binding-site motif enrichment analysis on predicted enhancers of H1 and H1 derived cell liens

To test whether the higher resolution of REPTILE enhancers improves TF-binding-site motif discovery, we conducted motif analysis on the REPTILE enhancer predictions in each human cell lineage. Homer (v4.8.3)[68] was used to identify the TF-binding-site motifs that were enriched in predicted enhancers in each human cell lineage. For each cell line, predicted enhancers were used as foreground (target) sequencers and Homer automatically selected the background sequencers (i.e. the default option). mm10 was used as the reference genome and we included the -nomotif option such that Homer only considered known motifs. In the next step, we selected motifs with q-values less than or equal to 0.05 as significantly enriched motifs. For each motif and predicted enhancers in each cell type, we calculated the degree of enrichment, which was defined as:

$$Enrichmentfoldchange = \frac{\%ofTargetSequenceswithMotif}{\%ofBackgroundSequenceswithMotif} \tag{3.5}$$

where Target Sequences refer to predicted enhancers and Background Sequencers are background regions automatically selected by Homer.

We asked whether this analysis could recapture motifs of the TFs known to function in that cell type. We downloaded the list of known transcription regulators for each human cell line from Xie et al.[47]. The mapping between TF names in the list and the motif names is available in Supplemental Table S4. We also conducted this analysis on the top 20,000 enhancer predictions from REPTILE and other existing methods in each human cell lineage. The results are shown in Figure 3.4.

Note that the lengths of enhancer predictions are different. As described in previous section, REPTILE enhancers have various lengths they have either the size of a DMR or 2kb (the length of sliding windows) depending how each of them was called as enhancer. Enhancer predictions from RFECS, DELTA and CSIANN are 2kb regions centered on the predicted enhancer centers. PEDLA made prediction on 200bp bins such that the size of PEDLA enhancers is 200bp.

3.11.17 Enhancer Prediction By Single Data type

To understand how informative single data type is, we used single epigenetic mark or only the open chromatin signature to predict the enhancer activity of regions in the test datasets (Figure 3.5A). We first calculated the enrichment score of an active mark (including open chromatin) or the depletion score of a repressive mark (mCG or H3K27me3) in tested regions. Then, we rank regions by their score and use AUPR to measure the how well the ranking distinguish active enhancers from negative regions. One common combination, DHS and H3K27ac, was also tested. The details of each approach are:

- **DHS**: The score of a tested region is the highest score of DHSs that overlap with it. If no overlapping DHS is found, its score is set to be negative infinity. The score of DHS corresponds to the signal value in the narrow peak format (https://genome.ucsc.edu/FAQ/FAQformat.html#format12).

- **DHS+H3K27ac**: The score of a tested region is the highest H3K27ac enrichment score in the DHSs that overlap with the tested region. If no DHS overlap is found, its score is set to be negative infinity. H3K27ac signal is the log2 RPM fold enrichment relative to control.

- **DNase-seq signal**: RPM of DNase-seq data for the tested region. The mean of values from replicates was used.

- **H3K27ac**: Average H3K27ac log2 RPM fold enrichment relative to control

- **mCG**: (1) x methylation level of tested region

- **DHS+mCG**: The score of a tested region is the largest negative CG methylation level in the DHSs that overlap with the tested region. If no DHS overlap is found, its score is set to be minus infinity.
- **H3K4me1**: H3K4me1 log2 RPM fold enrichment relative to control
- **H3K4me2**: H3K4me2 log2 RPM fold enrichment relative to control
- **H3K4me3**: H3K4me3 log2 RPM fold enrichment relative to control
- **H3K27me3**: (-1) x H3K27me3 log2 RPM fold enrichment relative to control
- **H3K9ac**: H3K9ac log2 RPM fold enrichment relative to control

3.11.18 Calling enhancers in human heart left ventricle

We specifically trained a human enhancer model for each method to generate enhancer prediction for human heart left ventricle because not all six previously used histone modifications are available in this tissue,. For PEDLA, DELTA, RFECS and CSIANN, this (re)training is almost identical to the previous training procedure but we limited the histone modifications to H3K4me1, H3K4me3, H3K27ac and H3K27me3, the histone marks that are available in both H1 and left ventricle. PEDLA also incorporated evolutionary conservation. The new enhancer models were retrained on the data of H1. REPTILE was trained on mCG data and histone modification data of H1, while left ventricle and H1 derived cells were used as reference. The DMR input for REPTILE was obtained by comparing the methylomes of left ventricle, H1 and H1 derived cells. In the prediction step, REPTILE used H1 and H1 derived cells as reference. Lastly, we applied these methods to generate enhancer predictions for left ventricle and the top 50,000 putative enhancers from each method were selected for later analyses.

3.11.19 Enrichment of disease-associated genetic variants in putative enhancers

To test for the enrichment of disease-associated SNPs in putative enhancers, we first downloaded the data of 5,654 non-coding GWAS SNPs from Maurano et al.[7]. The 5,654 SNPs were originally grouped into 15 categories based on the associated traits/diseases. We applied one-tail hypergeometric test to test the enrichment of SNPs from each category in the putative enhancers of left ventricle.

Specifically, for category c, the total number of SNPs in c is denoted as n_c and the total number of SNPs is $N = \sum_c n_c = 5,654$. Given a list of putative enhancers, the observed number of overlapped SNPs in category c is q_c and total observed number of overlapped SNPs is $Q = \sum_c q_c$. The p-value for SNPs in c is calculated as:

$$P(Kgeqq_c|N, n_c, Q, q_c) = \sum_{x=q_c}^{Q} \frac{\binom{n_c}{x}\binom{N-n_c}{Q-x}}{\binom{N}{Q}} \tag{3.6}$$

where K is a random variable representing the number of SNPs that are in category c and overlapped with putative enhancers. The fold enrichment of SNPs from category c in putative enhancers is calculated as $\frac{q_c/Q}{n_c/N}$.

Using the statistical test described above, for SNPs in each category, we tested for their enrichment in putative enhancers. We then used Benjamini-Hochberg approach to adjust p-values for multiple testing. P-value cutoff given 1%

false discovery rate (FDR) was used to call significant enrichment. This procedure was conducted separately for the putative enhancers from each method.

3.11.20 Linking REPTILE enhancers to target genes in left ventricle

To identify the target genes of REPTILE enhancers in left ventricle, we downloaded expression quantitative trait loci (eQTL) data of left ventricle from Genotype-Tissue Expression (GTEx) Project (version: v6p; file: Heart_Left_Ventricle_Analysis.v6p.signif_snpgene_pairs.txt.gz from GTEx_Analysis_v6p_eQTL.tar). The eQTLs that are within 2kb to any TSSs were filtered out. Then, we overlapped the REPTILE enhancers in left ventricle with the remaining eQTLs and assigned each putative enhancer to the gene linked to overlapping eQTL (if any). If multiple eQTLs are within one putative enhancer, the putative enhancer is assigned to all the genes linked to all eQTLs.

Next, based on the enhancer-gene assignment, we separated the genes that are linked to eQTLs into two groups. The first group consists of genes that are linked to at least one REPTILE enhancers, while the second group contains genes that are only linked to eQTLs outside of REPTILE enhancers. We then compared the expression levels of genes from these two groups. The gene expression data of left ventricle (from donor STL003) was obtained from Schultz et al.[19] and the expression level is represented in FPKM (fragments per kilobase of transcript per million mapped reads). Two-tailed Mann-Whitney test was conducted to test for the significance of difference in median expression levels of two groups of genes.

3.11.21 Testing the robustness of REPTILE given various input data

To test how sensitive REPTILEs performance is to various inputs, we ran REPTILE without DMRs (REPTILE w/o DMR), without reference epigenome (REPTILE w/o Ref) and with shuffled DMRs (REPTILE w/ shuf DMR) respectively. Since enhancer validation data is available for mouse samples, this test was done using mouse data. REPTILE w/o DMR performed prediction solely based on the epigenomic signature of query regions (e.g. 2kb sliding windows across the genome). REPTILE w/o Ref only uses the data of target sample (where prediction is generated) and does not calculate intensity deviation to describe tissue-specificity of epigenetic marks. Its enhancer model only uses the intensity of 7 marks as features. REPTILE w/ shuf DMR takes shuffled DMRs as input but its enhancer model is learned using unshuffled DMRs. We obtained the shuffled DMRs by shuffling the coordinates of DMRs within the genome while maintaining their lengths, which was done by using shuffleBed in bedtools[58].

REPTILE includes data of reference samples to capture the information in cell/tissue-specific epigenomic variation. However, it is unclear how the choice of reference would affect the prediction performance. To address this question, we implemented and benchmarked a different strategy of choosing reference. The new setup is called REPTILE alt Ref. In the new strategy, REPTILE always used mESCs, E11.5 Craniofacial and E11.5 Liver as reference samples for generating prediction. It is different from the original setup, which all mouse samples except target sample was used as reference. We applied this new setup to predict enhancers in E11.5 forebrain, midbrain, hindbrain, heart, limb and neural tube. For each target sample, the enhancer model was trained on data of mESCs using target sample, E11.5 Craniofacial and E11.5 Liver as reference, which corresponds to a scenario that only the data of the target sample and reference samples is available. The analysis of the prediction results is identical to the evaluation of the results from original setup.

3.12 Figures

Figure 3.1: **REPTILE improves enhancer identification by incorporating tissue-specific DNA methylation data.** **(A)** Differentially methylated regions (DMRs), typically smaller than query regions, serve as high-resolution enhancer candidates in overlapped query regions. **(B)** Example of a region (chr12: 29,660,800-29,668,600) where REPTILE uses base-resolution DNA methylation data to improve the resolution of enhancer prediction. Diagram of the gene model (GENCODE M2) in this region is shown at the top (Gene). DNA methylation displays mCG data of mESCs and eight E11.5 mouse tissues, where ticks represent methylated CG sites and their heights indicate the methylation level. Ticks on the forward strand are projected upward and ticks on the reverse strand are projected downward. Last track shows differentially methylated regions (DMRs) across these samples. Histone modification shows the log2 fold change of histone modification ChIP-seq data relative to input. Predictions from four computational methods are visualized in Enhancer prediction. Predictions from REPTILE best recapitulate the open chromatin data shown in DNase-seq. Light red rectangles mark the REPTILE putative enhancers, while the genomic locations of the midpoints (i.e. centers) are highlighted in red. **(C)** Workflow of REPTILE, including four major steps. 1) DMRs are identified by comparing the CG methylation profiles of target sample and the reference samples. 2) REPTILE integrates data in input files and represents query regions and DMRs as feature vectors (D). Yellow text on the top right corner shows the format for each input data type. 3) REPTILE trains an enhancer model based on the epigenomic signatures of known enhancers and negative sequences as well as the DMRs within them (red arrows). 4) Predictions are generated based on the enhancer model, DMR, query regions and epigenomic data (blue arrows). **(D)** Representation of one DMR or query region as a feature vector of intensity or intensity deviation of epigenetic marks. The 14 features used by REPTILE for the benchmark in this paper are shown. The -dev features in the vector are the intensity deviation features

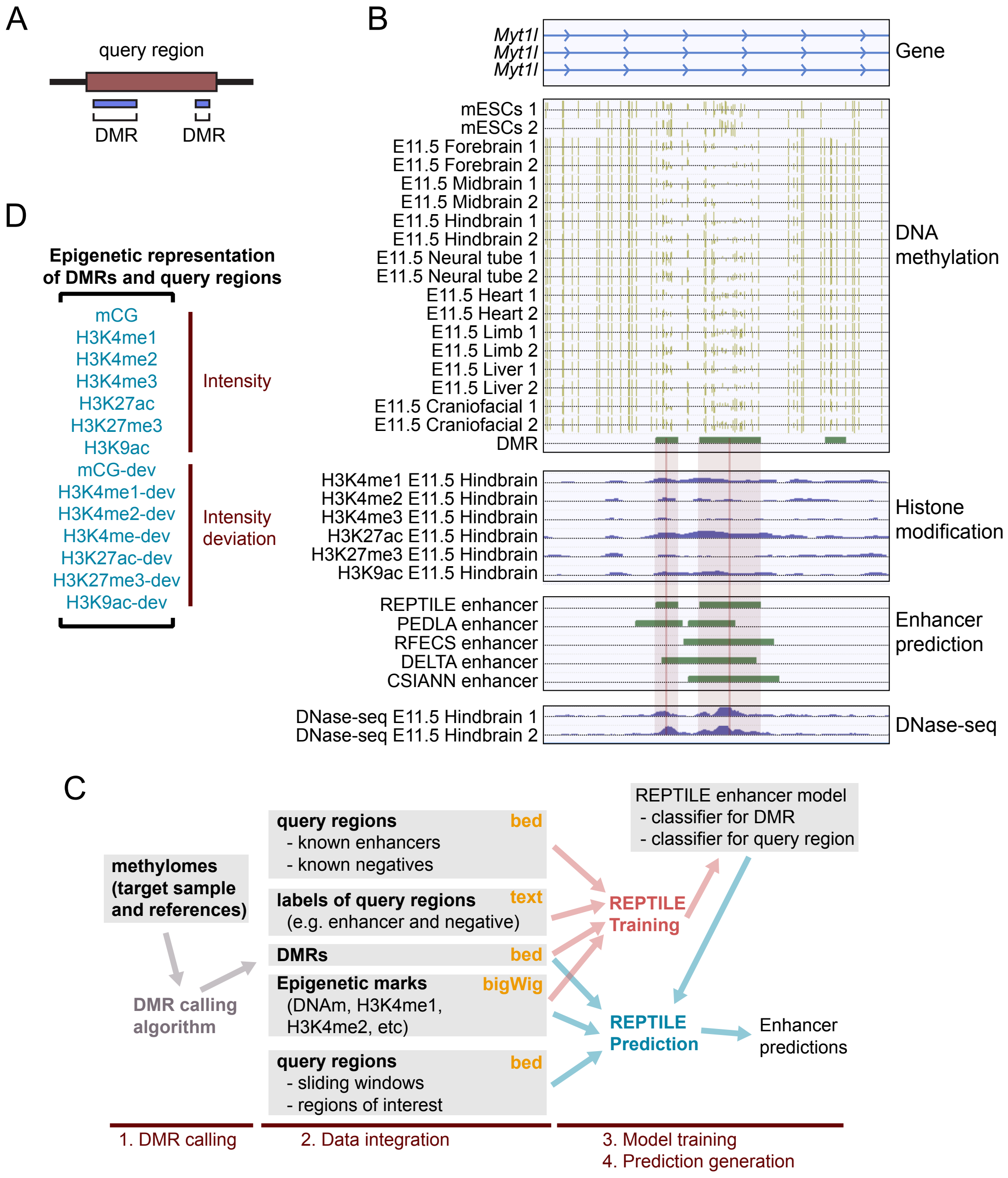
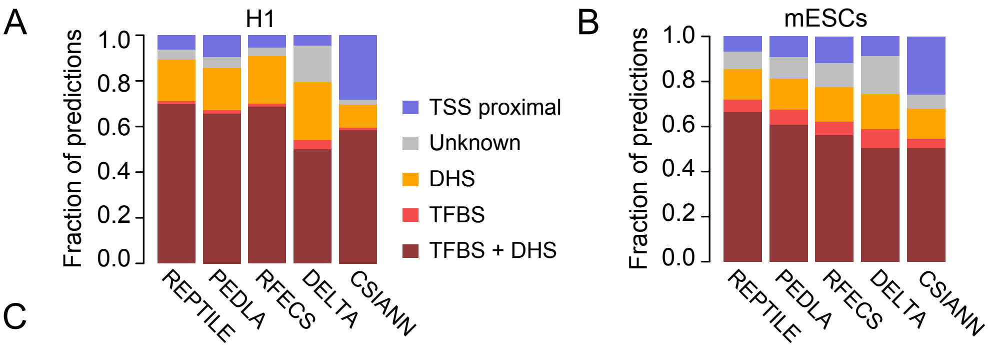

A
query region
DMR
DMR
B
Myt1l
Myt1l
Myt1l
Gene
mESCs 1
mESCs 2
E11.5 Forebrain 1
E11.5 Forebrain 2
E11.5 Midbrain 1
E11.5 Midbrain 2
E11.5 Hindbrain 1
E11.5 Hindbrain 2
E11.5 Neural tube 1
E11.5 Neural tube 2
E11.5 Heart 1
E11.5 Heart 2
E11.5 Limb 1
E11.5 Limb 2
E11.5 Liver 1
E11.5 Liver 2
E11.5 Craniofacial 1
E11.5 Craniofacial 2
DMR
DNA methylation
H3K4me1 E11.5 Hindbrain
H3K4me2 E11.5 Hindbrain
H3K4me3 E11.5 Hindbrain
H3K27ac E11.5 Hindbrain
H3K27me3 E11.5 Hindbrain
H3K9ac E11.5 Hindbrain
Histone modification
REPTILE enhancer
PEDLA enhancer
RFECS enhancer
DELTA enhancer
CSIANN enhancer
Enhancer prediction
DNase-seq E11.5 Hindbrain 1
DNase-seq E11.5 Hindbrain 2
DNase-seq
D
Epigenetic representation of DMRs and query regions
mCG
H3K4me1
H3K4me2
H3K4me3
H3K27ac
H3K27me3
H3K9ac
Intensity
mCG-dev
H3K4me1-dev
H3K4me2-dev
H3K4me-dev
H3K27ac-dev
H3K27me3-dev
H3K9ac-dev
Intensity deviation
C
methylomes (target sample and references)
DMR calling algorithm
query regions
- known enhancers
- known negatives
bed
labels of query regions
(e.g. enhancer and negative)
text
DMRs
bed
Epigenetic marks
(DNAm, H3K4me1, H3K4me2, etc)
bigWig
query regions
- sliding windows
- regions of interest
bed
REPTILE enhancer model
- classifier for DMR
- classifier for query region
REPTILE Training
REPTILE Prediction
Enhancer predictions
1. DMR calling
2. Data integration
3. Model training
4. Prediction generation

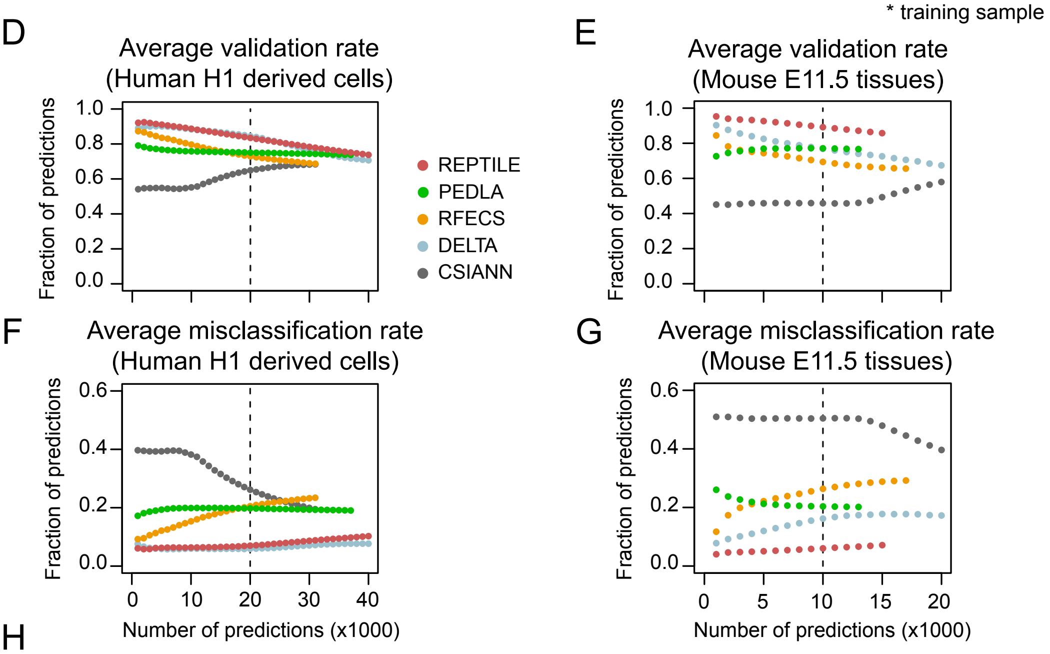
Figure 3.2: **REPTILE shows better enhancer prediction accuracy than existing methods.** (A-B) In H1 (A) and mESCs (B), the fractions of enhancers with their centers within 1kb to TFBS+DHS (dark red, both distal TFBSs and distal DHSs), TFBS (red, only distal TFBSs), DHS (orange, only distal DHSs), TSS proximal (overriding all other categories) or none of the above (grey, labeled as Unknown). Distal TFBS (DHSs) are defined as TFBSs (DHSs) that are at least 1kb away from any TSSs. TFBS, DHS and TFBS+DHS are considered as true positives, whereas TSS proximal is considered as false positive and misclassification. (C) Performances of all methods in eight test datasets that contain experimentally validated enhancers. Performances are measured by the area under precision-recall curve (AUPR). Best results in each test dataset are highlighted in red and second best results are marked in orange. The enhancer models used to make predictions in all samples were trained on data of mESCs. The baselines (AUPRs achieved using random guessing) for these datasets are shown in grey. Note that the AUPRs in different datasets cannot be compared because the fractions of validated enhancers are different. See Supplemental Fig. S1D for basic statistics of each dataset. (D-E) The validation rate of each method in human cell lines derived from H1 (D) and mouse tissues from E11.5 embryo (E), at different numbers of predictions. Validation rate is defined as the fraction of predictions whose centers are within 1kb from distal DHSs and are at least 1kb away from TSSs. (F-G) The misclassification rate of each method in human cell lines derived from H1 (F) and mouse tissues from E11.5 embryo (G). Misclassification rate is the fraction of predictions whose centers are within 1kb to TSSs. Vertical dash lines show the cutoffs used to get the final putative enhancer sets. (H) Examples of newly validated enhancers recapitulated by REPTILE enhancer predictions. Candidate enhancers were tested in transgenic mouse assays at E11.5. The enhancer name (mm or hs number), a representative transgenic embryo, and the tissues showing reproducible reporter gene expression (blue staining) are shown for each enhancer. mESCs - mouse embryonic stem cells; TFBS - transcription factor binding site; DHS - DNase hypersensitivity sites; TSS - transcription start site. See also Supplemental Methods for details.



C

	Area under Precision-Recall curve (AUPR)								Average Rank
	mESCs*	E11.5 Heart	E11.5 Limb	E11.5 Forebrain	E11.5 Midbrain	E11.5 Hindbrain	E11.5 Neural tube	Heart (Narlikar et al.)	
Baseline	0.62	0.20	0.13	0.13	0.11	0.07	0.06	0.39	
REPTILE	0.82	0.58	0.40	0.47	0.37	0.29	0.22	0.65	1.2
PEDLA	0.81	0.59	0.35	0.43	0.30	0.26	0.14	0.50	2.9
RFECS	0.80	0.54	0.33	0.40	0.35	0.30	0.19	0.50	2.9
DELTA	0.81	0.55	0.34	0.38	0.32	0.29	0.17	0.48	3.2
CSIANN	0.79	0.38	0.24	0.32	0.23	0.21	0.17	0.52	4.4

* training sample



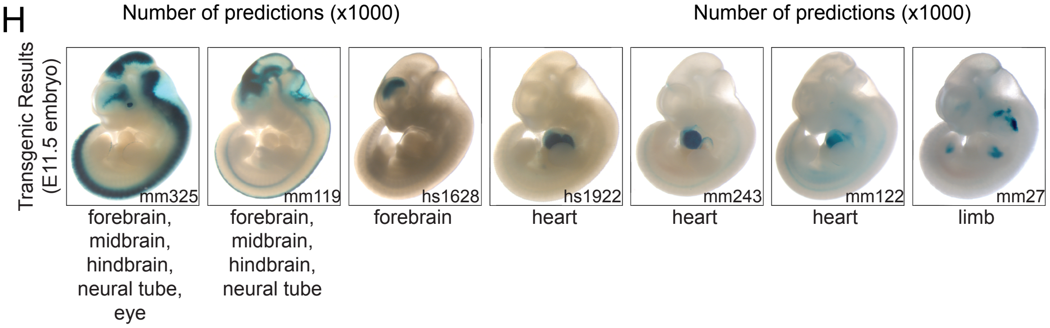


Figure 3.3: **The resolution of REPTILE predictions exceeds existing methods.** (A) Average distance between the centers of predictions and the closest distal DHSs in four human cell types derived from H1. Predictions whose centers are beyond 1kb away from the nearest distal DHS were considered as lack of support from open chromatin data and were not included in the calculation. Distal DHSs are at least 1kb away from any TSSs. (B) Average percentage of predictions whose centers are within 1kb to the closest distal DHS, in human cells derived from H1. (C) Average distance between the centers of predictions and the closest distal DHSs in mouse tissues from E11.5 embryo. (D) Average percentage of predictions whose centers are within 1kb to the closest distal DHS, in mouse tissues from E11.5 embryo. The metric value in each individual cell/tissue is shown as a point in the bar chart. MES - mesendoderm (MES); MSC - mesenchymal stem cells; NPC - neural progenitor cells; TRO - trophoblast-like cells; DHS - DNase hypersensitivity sites; TSS - transcription start site. See also Supplemental Methods for details.

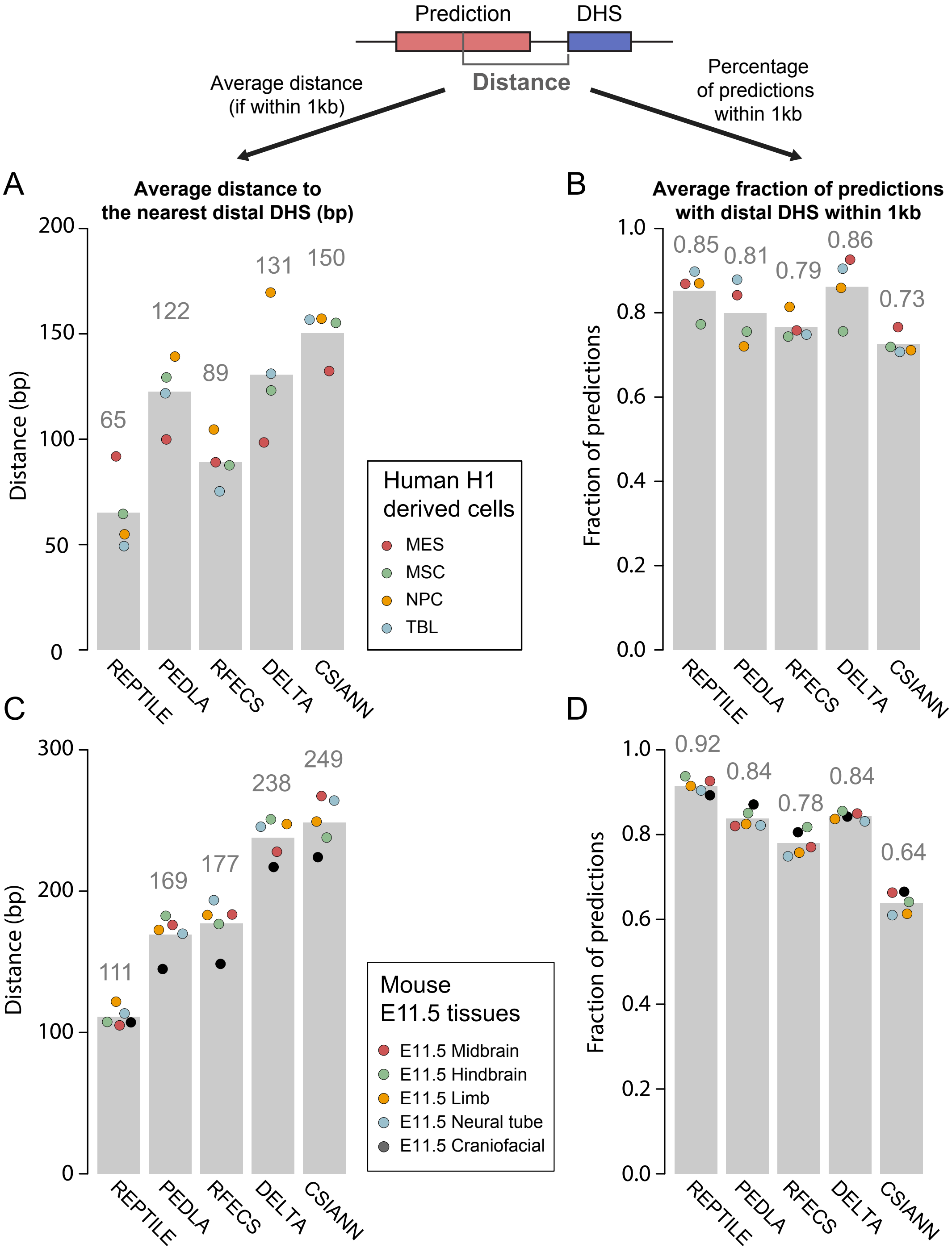

Prediction
DHS
Distance
Average distance (if within 1kb)
Percentage of predictions within 1kb
A
Average distance to the nearest distal DHS (bp)
Distance (bp)
0
50
100
150
200
65
122
89
131
150
REPTILE
PEDLA
RFECS
DELTA
CSIANN
Human H1 derived cells
● MES
● MSC
● NPC
● TBL
B
Average fraction of predictions with distal DHS within 1kb
Fraction of predictions
0.0
0.2
0.4
0.6
0.8
1.0
0.85
0.81
0.79
0.86
0.73
REPTILE
PEDLA
RFECS
DELTA
CSIANN
C
Distance (bp)
0
100
200
300
111
169
177
238
249
REPTILE
PEDLA
RFECS
DELTA
CSIANN
Mouse E11.5 tissues
● E11.5 Midbrain
● E11.5 Hindbrain
● E11.5 Limb
● E11.5 Neural tube
● E11.5 Craniofacial
D
Fraction of predictions
0.0
0.2
0.4
0.6
0.8
1.0
0.92
0.84
0.78
0.84
0.64
REPTILE
PEDLA
RFECS
DELTA
CSIANN

Figure 3.4: **REPTILE enhancers improve the detection of known transcriptional regulators for each cell type.** Enrichment of transcription-factor-binding-site motifs in the putative enhancers in H1 and H1 derived cells, respectively. Motif enrichments in each cell type were calculated on the predicted enhancers in matched cell type. Enrichment fold change is the fraction of predicted enhancers (target sequences) that contain a certain motif divided by the fraction of background sequences that contain the same motif. Highest enrichment of each motif in each cell type is marked in bold. Not significant enrichment (q-value ¿ 0.05) is shown in grey. The transcription factors (complex) listed under each cell type are known to function in that cell type, which were based on the list from Xie et al.[47]. See Supplemental Methods for details. H1 - H1 human embryonic stem cells; MES - mesendoderm (MES); MSC - mesenchymal stem cells; NPC - neural progenitor cells; TRO - trophoblast-like cells.

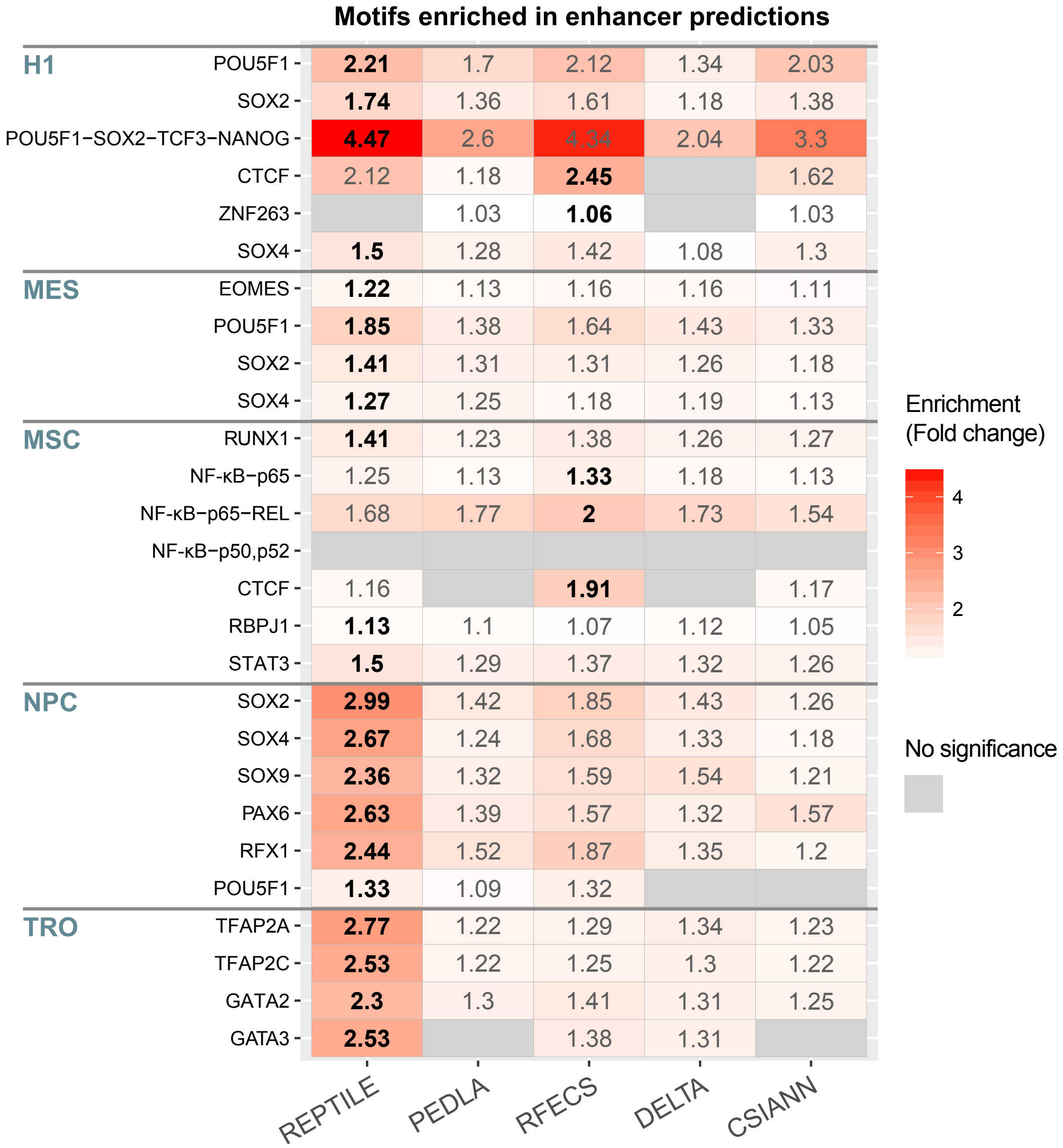

Motifs enriched in enhancer predictions
H1
POU5F1 2.21 1.7 2.12 1.34 2.03
SOX2 1.74 1.36 1.61 1.18 1.38
POU5F1−SOX2−TCF3−NANOG 4.47 2.6 4.34 2.04 3.3
CTCF 2.12 1.18 2.45 1.62
ZNF263 1.03 1.06 1.03
SOX4 1.5 1.28 1.42 1.08 1.3
MES
EOMES 1.22 1.13 1.16 1.16 1.11
POU5F1 1.85 1.38 1.64 1.43 1.33
SOX2 1.41 1.31 1.31 1.26 1.18
SOX4 1.27 1.25 1.18 1.19 1.13
MSC
RUNX1 1.41 1.23 1.38 1.26 1.27
NF-κB−p65 1.25 1.13 1.33 1.18 1.13
NF-κB−p65−REL 1.68 1.77 2 1.73 1.54
NF-κB−p50,p52
CTCF 1.16 1.91 1.17
RBPJ1 1.13 1.1 1.07 1.12 1.05
STAT3 1.5 1.29 1.37 1.32 1.26
NPC
SOX2 2.99 1.42 1.85 1.43 1.26
SOX4 2.67 1.24 1.68 1.33 1.18
SOX9 2.36 1.32 1.59 1.54 1.21
PAX6 2.63 1.39 1.57 1.32 1.57
RFX1 2.44 1.52 1.87 1.35 1.2
POU5F1 1.33 1.09 1.32
TRO
TFAP2A 2.77 1.22 1.29 1.34 1.23
TFAP2C 2.53 1.22 1.25 1.3 1.22
GATA2 2.3 1.3 1.41 1.31 1.25
GATA3 2.53 1.38 1.31
REPTILE
PEDLA
RFECS
DELTA
CSIANN
Enrichment
(Fold change)
4
3
2
No significance

Figure 3.5: **REPTILE enhancer confidence score is more predictive of enhancer activity than open chromatin or any single epigenetic mark.** (A) Performance of REPTILE and several enhancer prediction methods that are based on open chromatin, single epigenetic mark or the H3K27ac signal in open chromatin regions. The benchmark was done in four test datasets, where DNase-seq data is available in the corresponding samples. Performance is measured by the Area under Precision-Recall curve (AUPR). For each test dataset, the best performance(s) are highlighted in red and the second best are marked in orange. REPTILE generated scores on elements based on the enhancer model trained on data of mouse embryonic stem cells. DHS method assigned score to each element as the maximum normalized DNase-seq read count across all (1bp) overlapping DHSs. The score is 0 if the region contains no overlapping DHS. DHS+H3K27ac and DHS+mCG are similar to DHS but instead of DHS signal, it uses H3K27ac fold enrichment or CG methylation level as signal. The rest of the methods except mCG, DHS+mCG and H3K27me3 methods use the fold enrichment in whole elements as score. In contrast, mCG, DHS+mCG and H3K27me3 methods uses the signal values with reversed sign (i.e. depletion) because mCG and H3K27me3 are known to be repressive. (B-E) Precision of predicted enhancers that is based on the scores from REPTILE (red), DHS (orange), DHS+H3K27ac (light blue), DNase signal (grey) and H3K27ac (green) in E11.5 midbrain (B), hindbrain (C), neural tube (D) and limb (E). Precision is defined as the percentage of enhancer predictions that showed enhancer activity in vivo. DHS - DNase hypersensitivity sites; See also Supplemental Methods for details.

A

Method	AUPR			
	E11.5 Midbrain	E11.5 Hindbrain	E11.5 Neural tube	E11.5 Limb
REPTILE	**0.37**	**0.29**	**0.22**	**0.40**
DHS	0.22	0.20	**0.18**	**0.33**
DHS+H3K27ac	0.28	0.17	0.12	0.30
DNase-seq Signal	0.21	0.16	0.15	**0.33**
H3K27ac	0.26	0.18	0.12	0.31
mCG	0.20	0.14	0.10	0.23
DHS+mCG	**0.32**	**0.28**	0.17	0.32
H3K4me1	0.21	0.11	0.08	0.24
H3K4me2	0.18	0.13	0.09	0.19
H3K4me3	0.15	0.09	0.07	0.15
H3K27me3	0.08	0.07	0.06	0.15
H3K9ac	0.16	0.10	0.07	0.15

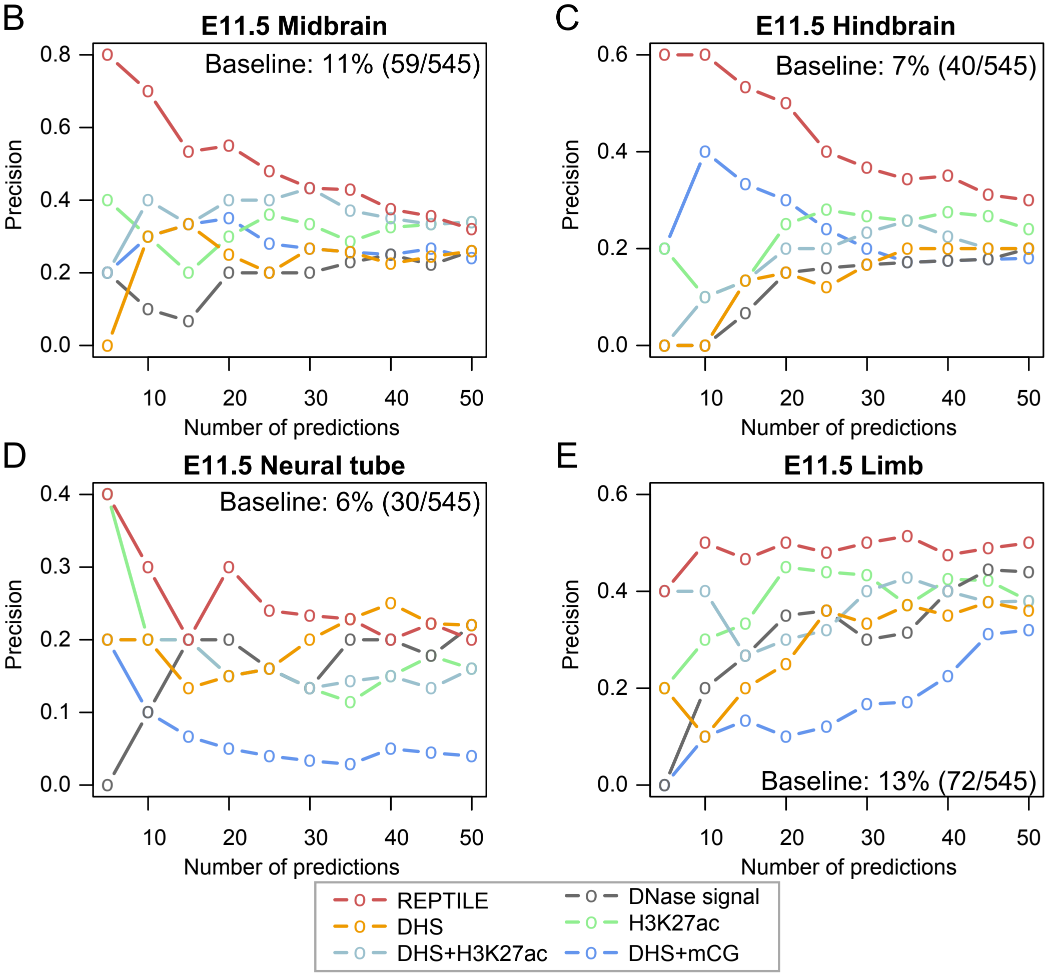
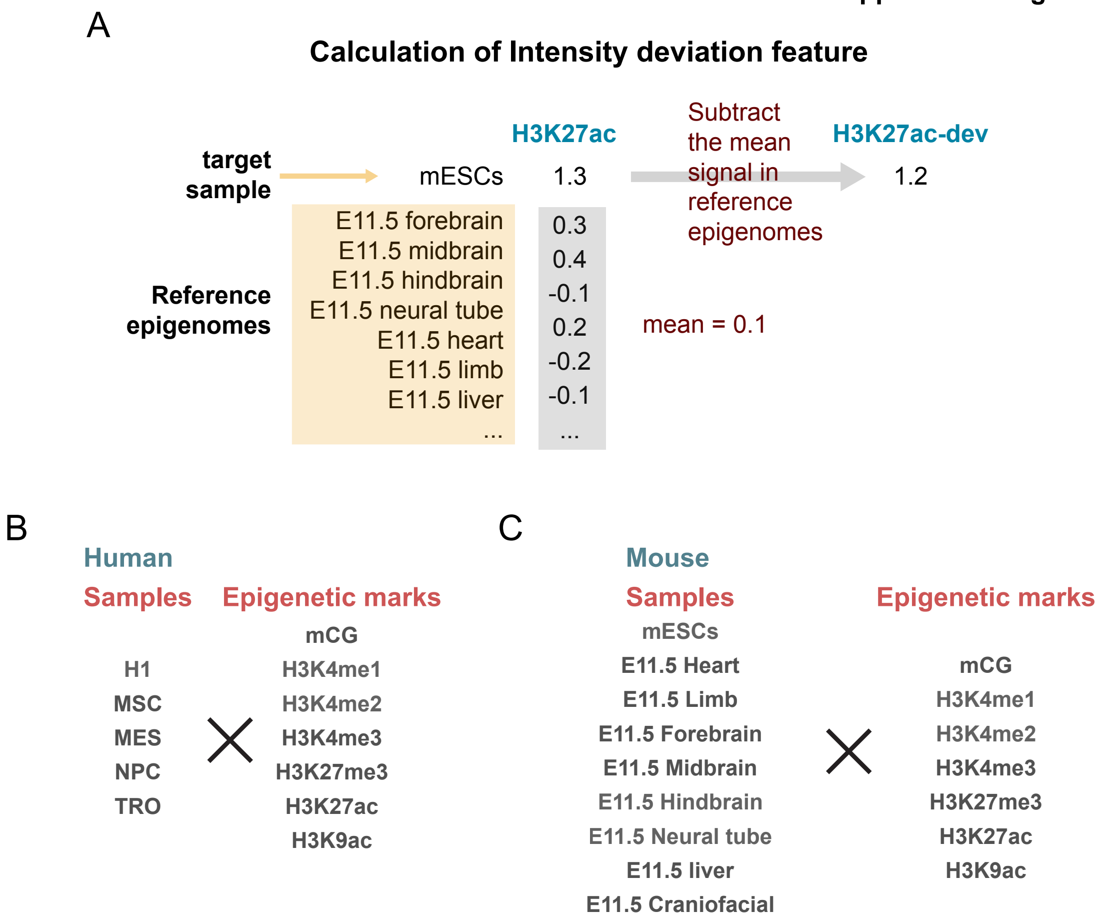


Figure 3.6: **Intensity deviation calculation and the enhancer validation and epigenomic data of various samples.** (A) An example of calculating intensity deviation. Given one epigenetic mark, the intensity in target sample (where predictions will be generated) is subtracted by the average intensity across reference samples. The result is the intensity deviation, which quantifies how the intensity in target sample is deviated from the default intensity (i.e. average value across reference samples). This feature captures the tissue/cell-specificity of epigenetic mark on a given region. (B-C) Human(B) and mouse (C) epigenomic data used in this paper, which includes all epigenetic marks (left) in all the samples (right). (D) Information about the eight enhancer validation datasets. The datasets were collected from Yue et al[32], VISTA enhancer browser[31] and Narlikar et al[49]. It also shows the basic statistics related to each dataset, including the total number of tested elements (Total), number of elements that showed evidence of enhancer activity (Positives) and its percentage out of all elements in the dataset (Positive%).

Supplemental Fig. S1



A

Calculation of Intensity deviation feature

H3K27ac

Subtract the mean signal in reference epigenomes

H3K27ac-dev

target sample

mESCs 1.3

1.2

Reference epigenomes

E11.5 forebrain 0.3

E11.5 midbrain 0.4

E11.5 hindbrain -0.1

E11.5 neural tube 0.2

E11.5 heart -0.2

E11.5 limb -0.1

E11.5 liver

... ...

mean = 0.1

B

Human

Samples: H1, MSC, MES, NPC, TRO

×

Epigenetic marks: mCG, H3K4me1, H3K4me2, H3K4me3, H3K27me3, H3K27ac, H3K9ac

C

Mouse

Samples: mESCs, E11.5 Heart, E11.5 Limb, E11.5 Forebrain, E11.5 Midbrain, E11.5 Hindbrain, E11.5 Neural tube, E11.5 liver, E11.5 Craniofacial

×

Epigenetic marks: mCG, H3K4me1, H3K4me2, H3K4me3, H3K27me3, H3K27ac, H3K9ac

D

Tissues	Source	Experiment	Total	Positives	Positive%
mESCs	Yue et al.	High-throughput reporter assay	211	131	62%
E11.5 Heart	VISTA	Transgenic reporter assay	545	110	20%
E11.5 Limb	VISTA	Transgenic reporter assay	545	72	13%
E11.5 Forebrain	VISTA	Transgenic reporter assay	545	70	13%
E11.5 Midbrain	VISTA	Transgenic reporter assay	545	59	11%
E11.5 Hindbrain	VISTA	Transgenic reporter assay	545	40	7%
E11.5 Neural tube	VISTA	Transgenic reporter assay	545	30	6%
Heart	Narlikar et al.	Transgenic reporter assay	36	14	39%

Figure 3.7: **Cross validation results and the evaluation of enhancer predictions by MERA data.** (A-B) 5-fold cross validation results on data of H1 (A) and mESCs (B). In each round of cross validation, we calculated a number of metrics to evaluate the performance of each method. These metrics include Accuracy, Precision, Recall (i.e. sensitivity), GM (geometric mean), F1 score, validation rate(s) and misclassification rate. For each method, the average rank is the mean of ranks in all metrics. Best rank is highlighted in red. (C) Percentage of MERA identified distal regulatory DNA elements that were recaptured by computational predictions. Bar chart shows the average percentage across all four MERA experiments for each method, while each circle shows the percentage in each MERA experiment. See also Supplemental Methods for details.

Supplemental Fig. S2

A

5-fold cross validation results on data in H1

			REPTILE	PEDLA	PEDLA (HMM)	RFECS	DELTA	CSIANN	ChromHMM	Segway
Number of prediction			20000	20000	20000	20000	20000	20000	20000	20000
Performance metrics	Accuracy		94.4%	94.0%	92.7%	94.3%	93.3%	92.7%	93.0%	90.6%
	Precision		74.1%	72.4%	63.6%	72.8%	68.7%	63.1%	72.1%	43.5%
	Recall/Sensitivity		58.9%	55.7%	45.0%	59.6%	48.2%	47.2%	36.8%	12.2%
	Specificity		97.9%	97.9%	97.5%	97.8%	97.8%	97.2%	98.6%	98.4%
	GM		75.9%	73.8%	66.2%	76.4%	68.7%	67.8%	60.2%	34.6%
	F1-score		65.6%	63.0%	52.7%	65.6%	56.7%	54.0%	48.7%	19.0%
	Validation rate	DHS	87.4%	84.4%	72.5%	89.4%	74.2%	68.5%	83.5%	74.6%
		TFBS	71.5%	67.3%	53.9%	70.1%	53.6%	59.6%	69.7%	53.2%
	Misclassification rate		7.3%	9.0%	10.1%	5.4%	4.7%	28.2%	6.8%	9.6%
Average Rank			**2.3**	3.5	6.2	**2.3**	4.5	6.2	4.6	6.6

B

5-fold cross validation results on data in mESCs

			REPTILE	PEDLA	PEDLA (HMM)	RFECS	DELTA	CSIANN	ChromHMM
Number of prediction			35000	35000	20516	35000	35000	35000	35000
Performance metrics	Accuracy		95.5%	95.5%	92.6%	95.1%	94.7%	93.3%	91.8%
	Precision		70.2%	70.2%	61.4%	66.4%	65.4%	59.4%	57.0%
	Recall/Sensitivity		88.3%	88.5%	52.3%	93.9%	88.6%	82.7%	41.1%
	Specificity		96.3%	96.3%	96.6%	95.3%	95.3%	94.3%	96.9%
	GM		92.2%	92.3%	71.1%	94.6%	91.9%	88.3%	63.1%
	F1-score		78.2%	78.3%	56.5%	77.8%	75.3%	69.1%	47.8%
	Validation	DHS	79.9%	74.9%	65.2%	71.6%	65.3%	63.8%	67.9%
		TFBS	71.8%	67.6%	57.2%	62.4%	58.6%	54.7%	60.5%
	Misclassification rate		7.6%	8.9%	15.8%	11.5%	8.6%	25.8%	5.8%
Average Rank			**2.3**	2.5	5.6	3.2	4.0	5.8	4.9

C

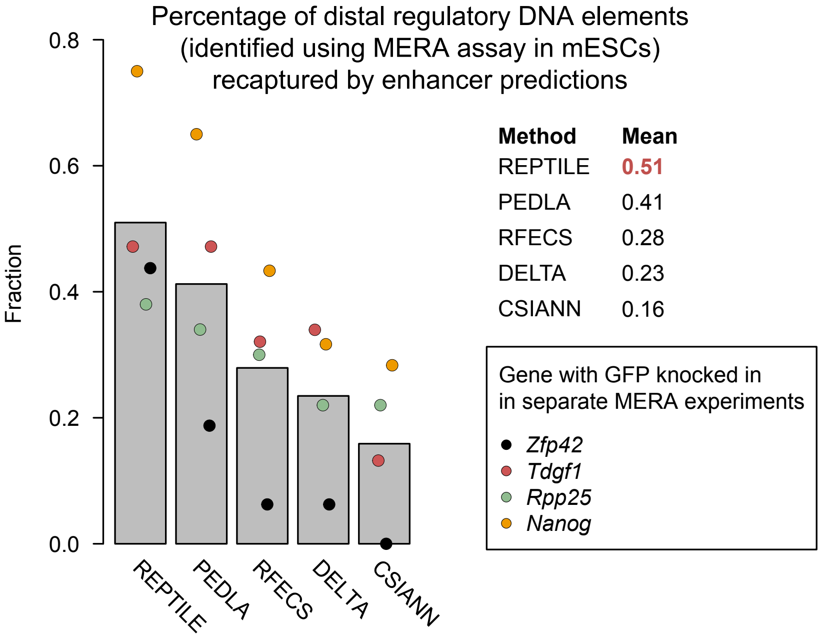


Figure 3.8: **Prediction accuracy and resolution in cells where the models were trained.** (A) The validation rate of each method in mESCs at different numbers of predictions. Validation rate is defined as the fraction of predictions whose centers are within 1kb from distal DHSs and are at least 1kb away from TSSs. (B) The misclassification rate of each method in mESCs. Misclassification rate is the fraction of predictions whose centers are within 1kb to TSSs. (C) The validation rate of each method in H1. (D) The misclassification rate of each method in H1. (E) Average distance between the centers of predictions and the closest distal DHSs in mESCs. Predictions whose centers are beyond 1kb away from the nearest distal DHS were considered as lack of support from open chromatin data and were not included in the calculation. Distal DHSs are at least 1kb away from any TSSs. (F) Average percentage of predictions whose centers are within 1kb to the closest distal DHS, in mESCs. (C) Average distance between the centers of predictions and the closest distal DHSs in H1. (D) Average percentage of predictions whose centers are within 1kb to the closest distal DHS, in mESCs. The metric value in each individual cell/tissue is shown as a point in the bar chart. mESC - mouse embryonic stem cell; DHS - DNase hypersensitivity sites; TSS - transcription start site. See also Supplemental Methods for details.

Supplemental Fig. S3

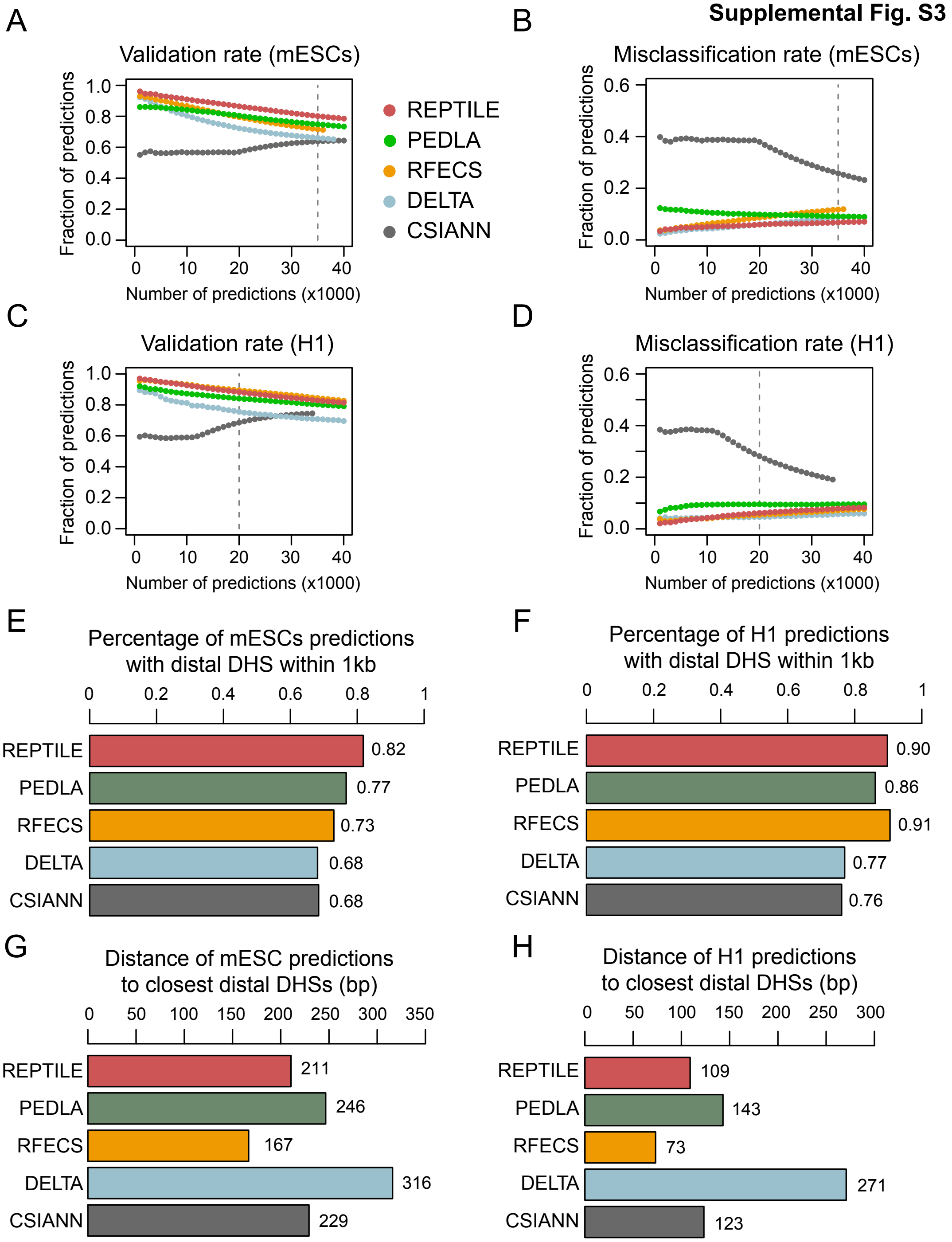


Figure 3.9: **Validation rate of enhancer predictions in human and mouse samples.** The validation rate of each method in different human and mouse cells and tissues at different numbers of predictions. Validation rate is defined as the fraction of predictions whose centers are within 1kb from distal DHSs and are at least 1kb away from TSSs. DHS - DNase hypersensitivity sites; TSS - transcription start site. See also Supplemental Methods for details.

Supplemental Fig. S4

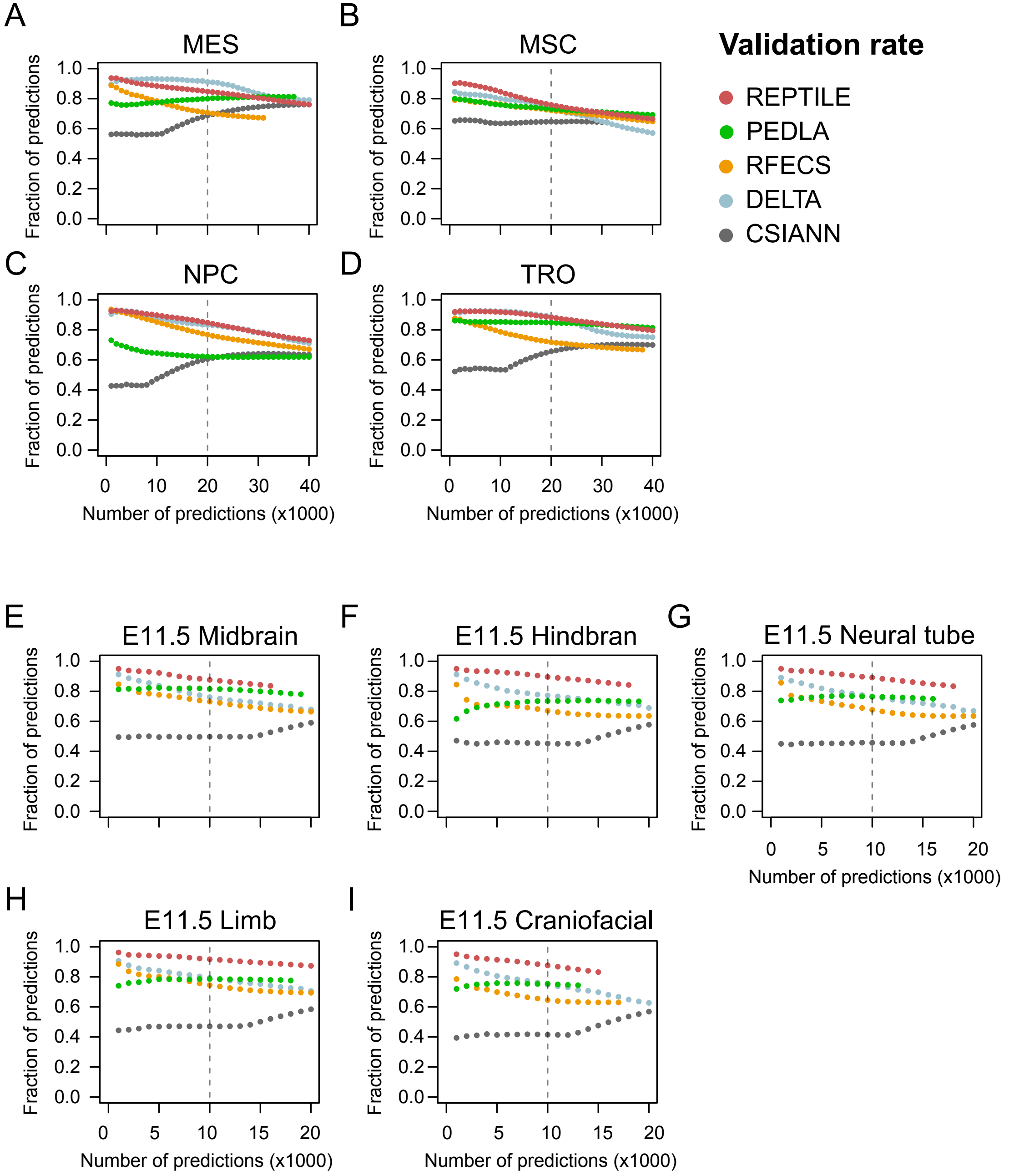


Figure 3.10: **Misclassification rate of enhancer predictions in human and mouse samples.** The misclassification rate of each method in different human and mouse cells and tissues at different numbers of predictions. Misclassification rate is the fraction of predictions whose centers are within 1kb to transcription start sites (TSSs). See also Supplemental Methods for details.

Supplemental Fig. S5

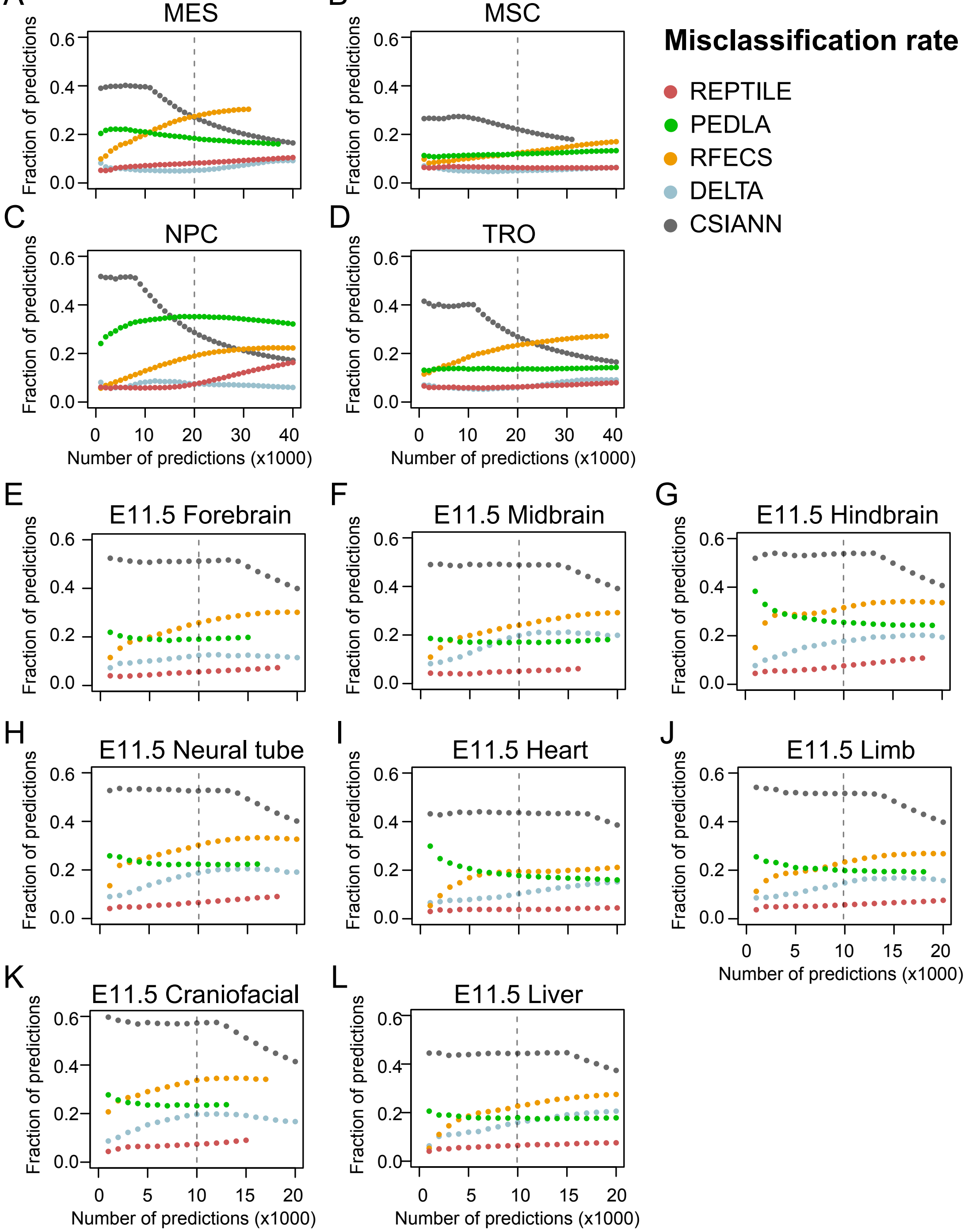
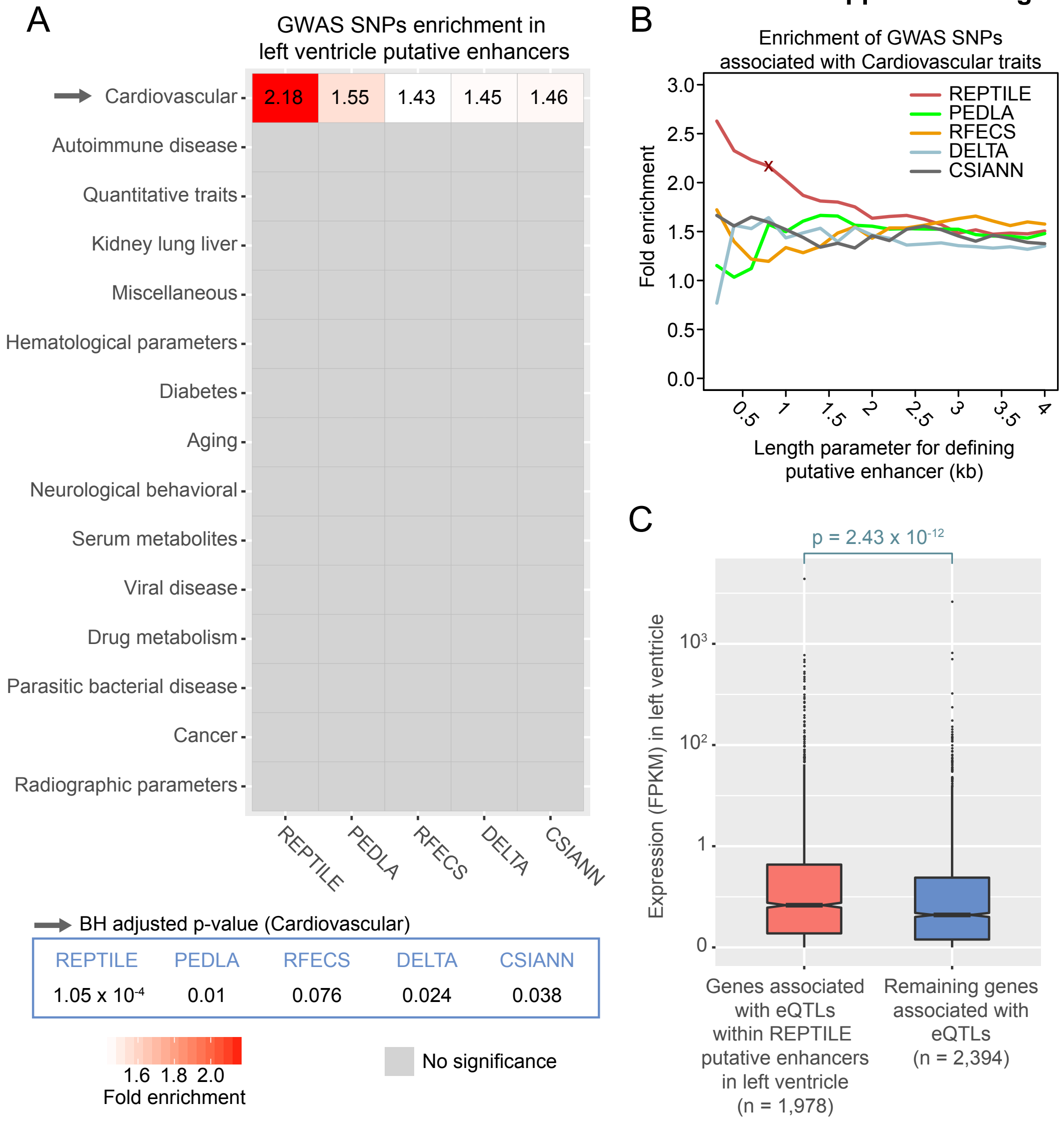


Figure 3.11: **REPTILE enhancers are enriched for non-coding GWAS SNPs and associated with increased expression of target genes.** (A) Enrichment of non-coding GWAS SNPs in putative enhancers of human heart left ventricle. Non-coding GWAS SNPs were categorized based the associated traits. One-tail hypergeometric test was used to test for significance. Benjamini-Hochberg approach was then used for multiple testing corrections. Non-coding GWAS SNPs and trait categories are from Maurano et al. [3]. See Supplemental Methods for more details. (B) Fold enrichment of the non-coding GWAS SNPs associated with Cardiovascular category in the left ventricle putative enhancers given different length parameters. The fold enrichment was calculated using putative enhancers that are defined as genomic regions of given length and centered at the predicted enhancer centers from each method. Red cross indicates the data point of REPTILE enhancers whose enhancer boundary is determined using DMR information. (C) In human heart left ventricle, genes associated with REPTILE enhancers show significantly higher expression than genes associated with other genomic loci. Expression quantitative trait loci (eQTLs) data was used to link REPTILE enhancers to the target genes. eQTLs that within 2kb to any transcription start sites were excluded for this analysis. Gene expression level is represented in FPKM (fragments per kilobase of transcript per million mapped reads).

Supplemental Fig. S6



BH adjusted p-value (Cardiovascular)

REPTILE	PEDLA	RFECS	DELTA	CSIANN
1.05 x 10^{-4}	0.01	0.076	0.024	0.038

Figure 3.12: **REPTILEs performance is robust to various subsets of input data.** (A) Performance of REPTILE models with various inputs and the four published methods in all test datasets. REPTILE w/o DMR has no DMR input. REPTILE w/o Ref makes predictions only based on the epigenomic data of target sample and thus does not include the intensity deviation features. REPTILE w/ shuf DMRs takes shuffled DMRs as input and makes prediction based on an enhancer model pre-trained in mESCs with DMRs before shuffling. For all methods, prediction results were generated by enhancer model trained in mouse embryonic stem cells. Model performance is measured by the Area under Precision-Recall curve (AUPR). For each test dataset, the best performance(s) are highlighted in red. The baselines are the AUPRs from random guessing for these datasets and are shown in grey. Note that the AUPRs in different datasets cannot be compared because the fractions of validated enhancers are different. See Supplemental Fig. S1D for the fraction in each dataset. (B-C) Bar charts showing the average distance from the center of mESCs enhancer predictions to nearest distal DHSs (B) and the fraction of predictions whose centers are within 1kb to distal DHSs (C). (D) Average distance between the centers of predictions and the nearest distal DHS. Predictions whose centers are beyond 1kb away from the nearest distal DHS were considered as lacking support from open chromatin data and were not included in the calculation. (E) Average percentage of predictions whose centers are within 1kb to the closest distal DHS. The results shown (E-F) are based on the enhancer predictions in E11.5 midbrain (red), hindbrain (green), limb (orange), neural tube (blue) and craniofacial (dark grey), where DNase-seq data was available. DHS - DNase hypersensitivity sites; See Supplemental Methods for details.

Supplemental Fig. S7

A

	AUPR							
	ESCs*	E11.5 Heart	E11.5 Limb	E11.5 Forebrain	E11.5 Midbrain	E11.5 Hindbrain	E11.5 Neural Tube	Heart (Narlikar et al.)
Baseline	0.62	0.20	0.13	0.13	0.11	0.07	0.06	0.39
REPTILE	**0.82**	0.58	0.40	0.47	0.37	0.29	**0.22**	0.65
REPTILE w/o DMR	0.79	0.58	**0.53**	**0.50**	**0.40**	**0.30**	0.19	0.62
REPTILE w/o Ref	0.79	0.48	0.41	0.46	0.28	0.29	0.20	**0.68**
REPTILE w/ shuf DMR	0.79	0.58	0.49	0.46	0.38	0.29	0.18	0.64
PEDLA	0.81	**0.59**	0.35	0.43	0.30	0.26	0.14	0.50
RFECS	0.80	0.54	0.33	0.40	0.35	0.30	0.19	0.50
DELTA	0.81	0.55	0.34	0.38	0.32	0.29	0.17	0.48
CSIANN	0.79	0.38	0.24	0.32	0.23	0.21	0.17	0.52

* training sample

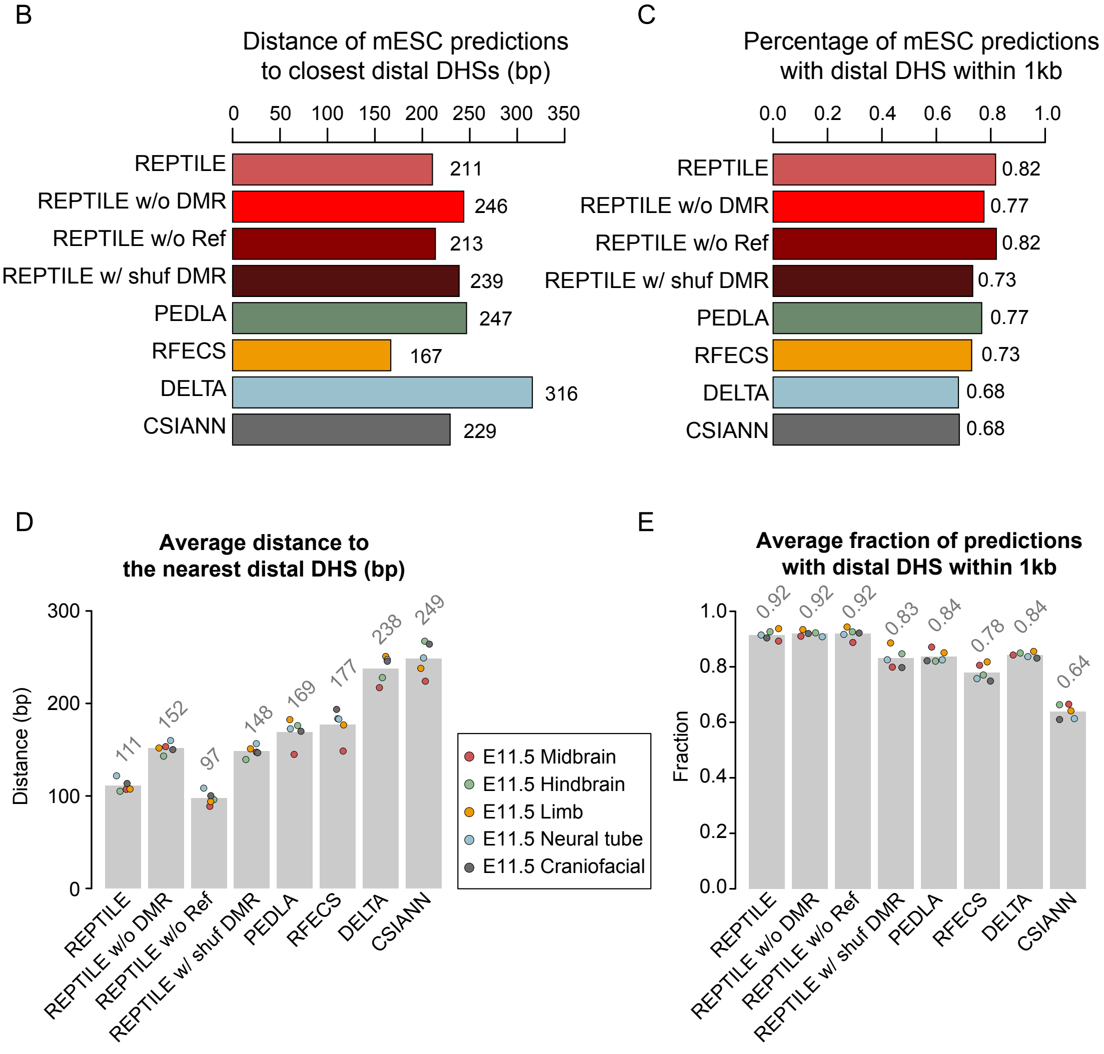


Figure 3.13: **REPTILEs performance is robust to the choice of reference.** (A) Performance of REPTILE models with different references and the four published methods in all test datasets. REPTILE alt Ref uses mESCs, E11.5 Craniofacial and E11.5 Liver as the reference in generating enhancer predictions. REPTILE w/o Ref does not use any reference. Model performance is measured by the Area under Precision-Recall curve (AUPR). The baselines are the AUPRs from random guessing for these datasets and are shown in grey. Note that the AUPRs in different datasets cannot be compared because the fractions of validated enhancers are different. (B) Average distance between the centers of predictions and the nearest distal DHS. Predictions whose centers are beyond 1kb away from the nearest distal DHS were considered as lacking support from open chromatin data and were not included in the calculation. (C) Average percentage of predictions whose centers are within 1kb to the closest distal DHS. The results shown (B-C) are based on the enhancer predictions in E11.5 midbrain (red), hindbrain (green), limb (orange) and neural tube (blue). DHS - DNase hypersensitivity sites; See Supplemental Methods for details.

Supplemental Fig. S8

A

	AUPR					
	E11.5 Heart	E11.5 Limb	E11.5 Forebrain	E11.5 Midbrain	E11.5 Hindbrain	E11.5 Neural Tube
Baseline	0.20	0.13	0.13	0.11	0.07	0.06
REPTILE	0.58	0.40	0.47	0.37	0.29	0.22
REPTILE - alt Ref *	0.61	0.46	0.47	0.35	0.35	0.21
REPTILE w/o Ref	0.48	0.41	0.46	0.28	0.29	0.20
PEDLA	0.59	0.35	0.43	0.30	0.26	0.14
RFECS	0.54	0.33	0.40	0.35	0.30	0.19
DELTA	0.55	0.34	0.38	0.32	0.29	0.17
CSIANN	0.38	0.24	0.32	0.23	0.21	0.17

* Only mESC, E11.5 Craniofacial and E11.5 Liver are used as reference

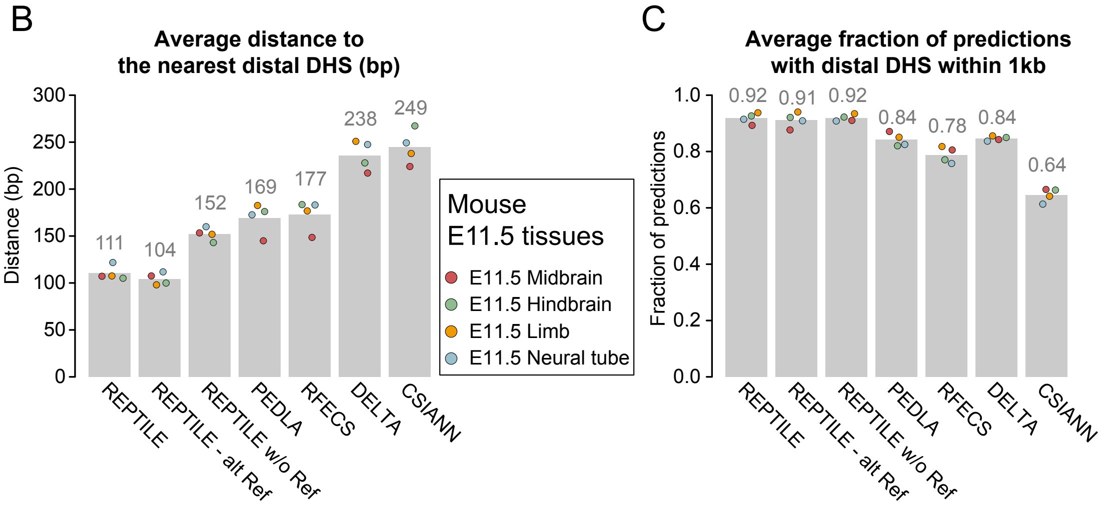
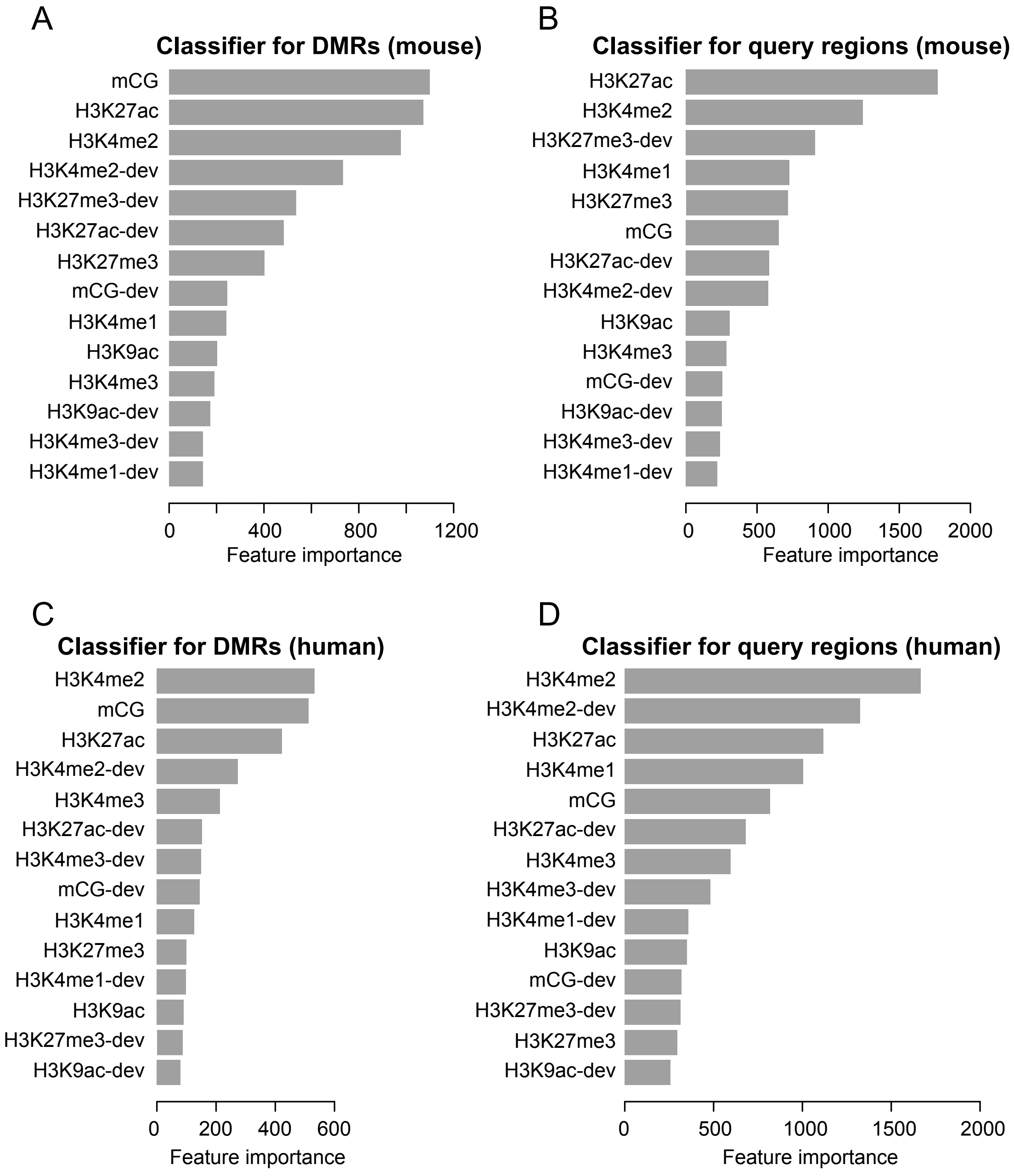


Figure 3.14: **Importance of features in the REPTILE enhancer model trained on mESCs.** (A-B) In the mouse REPTILE enhancer model, importance of features in the random forest classifiers for differentially methylated region (DMR) (A) and query regions (B). (C-D) In the human REPTILE enhancer model, importance of features in the random forest classifiers for (DMR) (C) and query regions (D). The importance is measured by the average decrease of Gini impurity index (meanDecreaseGini) from the randomForest R package[69].

Supplemental Fig. S9



3.13 References

[1] Laura A Lettice, Simon J H Heaney, Lorna A Purdie, Li Li, Philippe de Beer, Ben A Oostra, Debbie Goode, Greg Elgar, Robert E Hill, and Esther de Graaff. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human molecular genetics*, 12(14):1725–1735, jul 2003. ISSN 0964-6906 (Print).

[2] Tomoko Sagai, Masaki Hosoya, Youichi Mizushina, Masaru Tamura, and Toshihiko Shiroishi. Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. *Development (Cambridge, England)*, 132(4):797–803, feb 2005. ISSN 0950-1991 (Print). doi: 10.1242/dev.01613.

[3] Mark M Pomerantz, Nasim Ahmadiyeh, Li Jia, Paula Herman, Michael P Verzi, Harshavardhan Doddapaneni, Christine A Beckwith, Jennifer A Chan, Adam Hills, Matt Davis, Keluo Yao, Sarah M Kehoe, Heinz-Josef Lenz, Christopher A Haiman, Chunli Yan, Brian E Henderson, Baruch Frenkel, Jordi Barretina, Adam Bass, Josep Tabernero, Jose Baselga, Meredith M Regan, J Robert Manak, Ramesh Shivdasani, Gerhard A Coetzee, and Matthew L Freedman. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nature genetics*, 41(8): 882–884, aug 2009. ISSN 1546-1718 (Electronic). doi: 10.1038/ng.403.

[4] Olivier Harismendy, Dimple Notani, Xiaoyuan Song, Nazli G Rahim, Bogdan Tanasa, Nathaniel Heintzman, Bing Ren, Xiang-Dong Fu, Eric J Topol, Michael G Rosenfeld, and Kelly A Frazer. 9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response. *Nature*, 470(7333):264–268, feb 2011. ISSN 1476-4687 (Electronic). doi: 10.1038/nature09753.

[5] Dirk A Kleinjan and Veronica vanHeyningen. Long-Range Control of Gene Expression: Emerging Mechanisms and Disruption in Disease, jan 2005. ISSN 0002-9297 (Print).

[6] Noboru Jo Sakabe, Daniel Savic, and Marcelo A Nobrega. Transcriptional enhancers in development and disease. *Genome biology*, 13(1):238, 2012. ISSN 1474-760X (Electronic). doi: 10.1186/gb-2012-13-1-238.

[7] Matthew T. Maurano, Richard Humbert, Eric Rynes, Robert E. Thurman, Eric Haugen, Hao Wang, Alex P. Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, Anthony Shafer, Fidencio Neri, Kristen Lee, Tanya Kutyavin, Sandra Stehling-Sun, Audra K. Johnson, Theresa K. Canfield, Erika Giste, Morgan Diegel, Daniel Bates, R. Scott Hansen, Shane Neph, Peter J. Sabo, Shelly Heimfeld, Antony Raubitschek, Steven Ziegler, Chris Cotsapas, Nona Sotoodehnia, Ian Glass, Shamil R. Sunyaev, Rajinder Kaul, and John A. Stamatoyannopoulos. Systematic localization of common disease-associated variation in regulatory dna. *Science*, 337:1190–1195, Sep 2012.

[8] Yu Gyoung Tak and Peggy J. Farnham. Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenetics & Chromatin*, 8(1):57, 2015. ISSN 1756-8935. doi: 10.1186/s13072-015-0050-4. URL http://epigeneticsandchromatin.biomedcentral.com/articles/10.1186/s13072-015-0050-4.

[9] Menie Merika, Amy J Williams, Guoying Chen, Tucker Collins, and Dimitris Thanos. Recruitment of CBP/p300 by the IFNβ Enhanceosome Is Required for Synergistic Activation of Transcription. *Molecular Cell*, 1(2):277–287, jun 2016. ISSN 1097-2765. doi: 10.1016/S1097-2765(00)80028-3. URL http://dx.doi.org/10.1016/S1097-2765(00)80028-3.

[10] Nathaniel D Heintzman, Rhona K Stuart, Gary Hon, Yutao Fu, Christina W Ching, R David Hawkins, Leah O Barrera, Sara Van Calcar, Chunxu Qu, Keith a Ching, Wei Wang, Zhiping Weng, Roland D Green, Gregory E Crawford, and Bing Ren. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics*, 39(3):311–318, 2007. ISSN 1061-4036. doi: 10.1038/ng1966.

[11] Nathaniel D Heintzman, Gary C Hon, R David Hawkins, Pouya Kheradpour, Alexander Stark, Lindsey F Harp, Zhen Ye, Leonard K Lee, Rhona K Stuart, Christina W Ching, Keith a Ching, Jessica E Antosiewicz, Hui Liu, Xinmin Zhang, Roland D Green, Ron Stewart, James a Thomson, and Gregory E Crawford. Histone modification at human enhancers reflect global cell-type specific gene expression. *Nature*, 459(7243):108–112, 2009. doi: 10.1038/nature07829.Histone.

[12] Menno P Creyghton, Albert W Cheng, G Grant Welstead, Tristan Kooistra,

Bryce W Carey, Eveline J Steine, Jacob Hanna, Michael A Lodato, Garrett M Frampton, Phillip A Sharp, Laurie A Boyer, Richard A Young, and Rudolf Jaenisch. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, 107(50):21931–21936, 2010. doi: 10.1073/pnas.1016071107. URL http://www.pnas.org/content/107/50/21931.abstract.

[13] Dimitrios Kleftogiannis, Panos Kalnis, and Vladimir B Bajic. Progress and challenges in bioinformatics approaches for enhancer identification. *Briefings in bioinformatics*, (August):bbv101–, 2015. ISSN 1477-4054. doi: 10.1093/bib/bbv101. URL http://bib.oxfordjournals.org/content/early/2015/12/03/bib.bbv101.long.

[14] A. Bird. Dna methylation patterns and epigenetic memory. *Genes Dev*, 16:6, 2002.

[15] J. A. Law and S. E. Jacobsen. Establishing, maintaining and modifying dna methylation patterns in plants and animals. *Nat Rev Genet*, 11:204–220, Mar 2010.

[16] Peter a. Jones. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13(7):484–492, 2012. ISSN 1471-0056. doi: 10.1038/nrg3230. URL http://dx.doi.org/10.1038/nrg3230.

[17] Zachary D. Smith and Alexander Meissner. Dna methylation: roles in mammalian development. Feb 2013.

[18] R. Lister, M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, J. R. Nery, L. Lee, Z. Ye, Q. M. Ngo, L. Edsall, J. Antosiewicz-Bourget, R. Stewart, V. Ruotti, A. H. Millar, J. A. Thomson, B. Ren, and J. R. Ecker. Human dna methylomes at base resolution show widespread epigenomic differences. *Nature*, 462:315–322, 2009.

[19] M. D. Schultz, Y. He, J. W. Whitaker, M. Hariharan, E. A. Mukamel, D. Leung, N. Rajagopal, J. R. Nery, M. A. Urich, H. Chen, S. Lin, Y. Lin, I. Jung, A. D. Schmitt, S. Selvaraj, B. Ren, T. J. Sejnowski, W. Wang, and J. R. Ecker. Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature*, 523(7559):212–216, Jul 2015.

[20] Michael J. Ziller, Hongcang Gu, Fabian Müller, Julie Donaghey, Linus T-Y Tsai, Oliver Kohlbacher, Philip L. De Jager, Evan D. Rosen, David A. Bennett, Bradley E. Bernstein, Andreas Gnirke, and Alexander Meissner. Charting a dynamic dna methylation landscape of the human genome. *Nature*, 500:477–481, Aug 2013.

[21] Katherine E. Varley, Jason Gertz, Kevin M. Bowling, Stephanie L. Parker, Timothy E. Reddy, Florencia Pauli-Behn, Marie K. Cross, Brian A. Williams, John A. Stamatoyannopoulos, Gregory E. Crawford, Devin M. Absher, Barbara J. Wold, and Richard M. Myers. Dynamic dna methylation across diverse human cell lines and tissues. *Genome Res*, 23:555–567, Mar 2013.

[22] Y. He and J. R. Ecker. Non-CG Methylation in the Human Genome. *Annu Rev Genomics Hum Genet*, 16:55–77, 2015.

[23] M. B. Stadler, R. Murr, L. Burger, R. Ivanek, F. Lienert, A. Schöler, C. Wirbelauer, E. J. Oakeley, D. Gaidatzis, V. K. Tiwari, and D. Schübeler. Dna-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, 480:490–495, Dec 2011.

[24] Syed Khund Sayeed, Jianfei Zhao, Bangalore K. Sathyanarayana, Jaya Prakash Golla, and Charles Vinson. C/EBPβ (CEBPB) protein binding to the C/EBP—CRE DNA 8-mer TTGC—GTCA is inhibited by 5hmC and enhanced by 5mC, 5fC, and 5caC in the CG dinucleotide. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, 1849(6):583–589, 2015. ISSN 18749399. doi: 10.1016/j.bbagrm.2015.03.002. URL http://linkinghub.elsevier.com/retrieve/pii/S187493991500067X.

[25] R. C. O'Malley, S. S. Huang, L. Song, M. G. Lewsey, A. Bartlett, J. R. Nery, M. Galli, A. Gallavotti, and J. R. Ecker. Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell*, 165(5):1280–1292, May 2016.

[26] Dominique C. Stephens and GregoryM.K. Poon. Differential sensitivity to methylated DNA by ETS-family transcription factors is intrinsically encoded in their DNA-binding domains. *Nucleic Acids Research*, page gkw528, 2016. ISSN 0305-1048. doi: 10.1093/nar/gkw528. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkw528.

[27] T. Xu, B. Li, M. Zhao, K. E. Szulwach, R. C. Street, L. Lin, B. Yao, F. Zhang, P. Jin, H. Wu, and Z. S. Qin. Base-resolution methylation patterns accurately predict transcription factor bindings in vivo. *Nucleic Acids Research*, pages 1–10, 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv151. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkv151.

[28] Woochang Hwang, Verity F Oliver, Shannath L Merbs, Heng Zhu, and Jiang Qian. Prediction of promoters and enhancers using multiple DNA methylation-associated features. *BMC Genomics*, 16(7):1–13, 2015. ISSN 1471-2164. doi: 10.1186/1471-2164-16-S7-S11. URL http://dx.doi.org/10.1186/1471-2164-16-S7-S11.

[29] Peter J. Park. ChIP-Seq: advantages and challenges of a maturing technology. *Nature reviews. Genetics*, 10(10):669–680, 2009. ISSN 1471-0056. doi: 10.1038/nrg2641. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3191340/.

[30] Gary C. Hon, Nisha Rajagopal, Yin Shen, David F. McCleary, Feng Yue, My D. Dang, and Bing Ren. Epigenetic memory at embryonic enhancers identified in dna methylation maps from adult mouse tissues. *Nat Genet*, 45: 1198–1206, Oct 2013.

[31] Axel Visel, Simon Minovitsky, Inna Dubchak, and Len a. Pennacchio. VISTA Enhancer Browser - A database of tissue-specific human enhancers. *Nucleic Acids Research*, 35(SUPPL. 1):88–92, 2007. ISSN 03051048. doi: 10.1093/nar/gkl822.

[32] F Yue, Y Cheng, A Breschi, J Vierstra, W Wu, T Ryba, R Sandstrom, Z Ma, C Davis, B D Pope, Y Shen, D D Pervouchine, S Djebali, R E Thurman, R Kaul, E Rynes, A Kirilusha, G K Marinov, B A Williams, D Trout, H Amrhein, K Fisher-Aylor, I Antoshechkin, G DeSalvo, L H See, M Fastuca, J Drenkow, C Zaleski, A Dobin, P Prieto, J Lagarde, G Bussotti, A Tanzer, O Denas, K Li, M A Bender, M Zhang, R Byron, M T Groudine, D McCleary, L Pham, Z Ye, S Kuan, L Edsall, Y C Wu, M D Rasmussen, M S Bansal, M Kellis, C A Keller, C S Morrissey, T Mishra, D Jain, N Dogan, R S Harris, P Cayting, T Kawli, A P Boyle, G Euskirchen, A Kundaje, S Lin, Y Lin, C Jansen, V S Malladi, M S Cline, D T Erickson, V M Kirkup, K Learned, C A Sloan, K R Rosenbloom, B Lacerda de Sousa, K Beal, M Pignatelli, P Flicek, J Lian, T Kahveci, D Lee, W J Kent, M Ramalho

Santos, J Herrero, C Notredame, A Johnson, S Vong, K Lee, D Bates, F Neri, M Diegel, T Canfield, P J Sabo, M S Wilken, T A Reh, E Giste, A Shafer, T Kutyavin, E Haugen, D Dunn, A P Reynolds, S Neph, R Humbert, R S Hansen, M De Bruijn, L Selleri, A Rudensky, S Josefowicz, R Samstein, E E Eichler, S H Orkin, D Levasseur, T Papayannopoulou, K H Chang, A Skoultchi, S Gosh, C Disteche, P Treuting, Y Wang, M J Weiss, G A Blobel, X Cao, S Zhong, T Wang, P J Good, R F Lowdon, L B Adams, X Q Zhou, M J Pazin, E A Feingold, B Wold, J Taylor, A Mortazavi, S M Weissman, J A Stamatoyannopoulos, M P Snyder, R Guigo, T R Gingeras, D M Gilbert, R C Hardison, M A Beer, B Ren, and Encode Consortium Mouse. A comparative encyclopedia of DNA elements in the mouse genome. *Nature*, 515(7527):355–364, 2014. ISSN 0028-0836. doi: 10.1038/nature13992. URL http://www.ncbi.nlm.nih.gov/pubmed/25409824{%}5Cnhttp://www.nature.com/nature/journal/v515/n7527/pdf/nature13992.pdf.

[33] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324. URL http://dx.doi.org/10.1023/A:1010933404324.

[34] Feng Liu, Hao Li, Chao Ren, Xiaochen Bo, and Wenjie Shu. PEDLA: predicting enhancers with a deep learning-based algorithmic framework. *Scientific Reports*, 6:28517, jun 2016. URL http://dx.doi.org/10.1038/srep28517http://10.0.4.14/srep28517http://www.nature.com/articles/srep28517{#}supplementary-information.

[35] Nisha Rajagopal, Wei Xie, Yan Li, Uli Wagner, Wei Wang, John Stamatoyannopoulos, Jason Ernst, Manolis Kellis, and Bing Ren. RFECS: A Random-Forest Based Algorithm for Enhancer Identification from Chromatin State. *PLoS Computational Biology*, 9(3), 2013. ISSN 1553734X. doi: 10.1371/journal.pcbi.1002968.

[36] Yiming Lu, Wubin Qu, Guangyu Shan, and Chenggang Zhang. DELTA: A Distal Enhancer Locating Tool Based on AdaBoost Algorithm and Shape Features of Chromatin Modifications. *Plos One*, 10(6):e0130622, 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0130622. URL http://dx.plos.org/10.1371/journal.pone.0130622.

[37] Hiram A. Firpi, Duygu Ucar, and Kai Tan. Discover regulatory DNA elements using chromatin signatures and artificial neural network. *Bioinformatics*, 26

(13):1579–1586, 2010. ISSN 13674803. doi: 10.1093/bioinformatics/btq248.

[38] Nisha Rajagopal, Sharanya Srinivasan, Kameron Kooshesh, Yuchun Guo, Matthew D Edwards, Budhaditya Banerjee, Tahin Syed, Bart J M Emons, David K Gifford, and Richard I Sherwood. High-throughput mapping of regulatory DNA. *Nat Biotech*, 34(2):167–174, feb 2016. ISSN 1087-0156. URL http://dx.doi.org/10.1038/nbt.3468http://10.0.4.14/nbt.3468http://www.nature.com/nbt/journal/v34/n2/abs/nbt.3468.html{#}supplementary-information.

[39] Alan P. Boyle, Sean Davis, Hennady P. Shulha, Paul Meltzer, Elliott H. Margulies, Zhiping Weng, Terrence S. Furey, and Gregory E. Crawford. High-Resolution Mapping andCharacterization of Open Chromatin across the Genome. *Cell*, 132(2):311–322, 2008. ISSN 00928674. doi: 10.1016/j.cell.2007.12.014.

[40] Jason D Buenrostro, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods*, 10(12):1213–8, 2013. ISSN 1548-7105. doi: 10.1038/nmeth.2688. URL http://dx.doi.org/10.1038/nmeth.2688.

[41] Lijing Yao, Hui Shen, Peter W Laird, Peggy J Farnham, and Benjamin P Berman. Inferring regulatory element landscapes and transcription factor networks from cancer methylomes. *Genome biology*, 16(105):1–21, 2015. ISSN 1465-6914. doi: 10.1186/s13059-015-0668-3. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4460959{&}tool=pmcentrez{&}rendertype=abstract.

[42] Suhn Kyong Rhie, Yu Guo, Yu Gyoung Tak, Lijing Yao, Hui Shen, Gerhard A. Coetzee, Peter W. Laird, and Peggy J. Farnham. Identification of activated enhancers and linked transcription factors in breast, prostate, and kidney tumors by tracing enhancer networks using epigenetic traits. *Epigenetics & Chromatin*, 9(1):50, 2016. ISSN 1756-8935. doi: 10.1186/s13072-016-0102-4. URL http://epigeneticsandchromatin.biomedcentral.com/articles/10.1186/s13072-016-0102-4.

[43] Genevieve D Erwin, Nir Oksenberg, Rebecca M Truty, Dennis Kostka, Karl K Murphy, Nadav Ahituv, Katherine S Pollard, and John A Capra. Integrat-

ing Diverse Datasets Improves Developmental Enhancer Prediction. *PLoS Comput Biol*, 10(6):1–20, 2014. doi: 10.1371/journal.pcbi.1003677. URL http://dx.doi.org/10.1371{%}2Fjournal.pcbi.1003677.

[44] E Birney, J A Stamatoyannopoulos, A Dutta, R Guigo, T R Gingeras, E H Margulies, Z Weng, M Snyder, E T Dermitzakis, and R E Thurman. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447, 2007. doi: 10.1038/nature05874. URL http://dx.doi.org/10.1038/nature05874.

[45] ENCODE Project Consortium, Bradley E Bernstein, Ewan Birney, Ian Dunham, Eric D Green, Chris Gunter, and Michael Snyder. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012. ISSN 1476-4687. doi: nature11247[pii] \n10.1038/nature11247. URL http://www.nature.com/doifinder/10.1038/nature11247{%}5Cnpapers3://publication/doi/10.1038/nature11247.

[46] Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J. Ziller, Viren Amin, John W. Whitaker, Matthew D. Schultz, Lucas D. Ward, Abhishek Sarkar, Gerald Quon, Richard S. Sandstrom, Matthew L. Eaton, Yi-Chieh Wu, Andreas R. Pfenning, Xinchen Wang, Melina Claussnitzer, Yaping Liu, Cristian Coarfa, R. Alan Harris, Noam Shoresh, Charles B. Epstein, Elizabeta Gjoneska, Danny Leung, Wei Xie, R. David Hawkins, Ryan Lister, Chibo Hong, Philippe Gascard, Andrew J. Mungall, Richard Moore, Eric Chuah, Angela Tam, Theresa K. Canfield, R. Scott Hansen, Rajinder Kaul, Peter J. Sabo, Mukul S. Bansal, Annaick Carles, Jesse R. Dixon, Kai-How Farh, Soheil Feizi, Rosa Karlic, Ah-Ram Kim, Ashwinikumar Kulkarni, Daofeng Li, Rebecca Lowdon, GiNell Elliott, Tim R. Mercer, Shane J. Neph, Vitor Onuchic, Paz Polak, Nisha Rajagopal, Pradipta Ray, Richard C. Sallari, Kyle T. Siebenthall, Nicholas A. Sinnott-Armstrong, Michael Stevens, Robert E. Thurman, Jie Wu, Bo Zhang, Xin Zhou, Arthur E. Beaudet, Laurie A. Boyer, Philip L. De Jager, Peggy J. Farnham, Susan J. Fisher, David Haussler, Steven J. M. Jones, Wei Li, Marco A. Marra, Michael T. McManus, Shamil Sunyaev, James A. Thomson, Thea D. Tlsty, Li-Huei Tsai, Wei Wang, Robert A. Waterland, Michael Q. Zhang, Lisa H. Chadwick, Bradley E. Bernstein, Joseph F. Costello, Joseph R. Ecker, Martin Hirst, Alexander Meissner, Aleksandar Milosavljevic, Bing Ren, John A. Stamatoyannopoulos, Ting Wang, Manolis Kellis, Andreas Pfenning, Melina ClaussnitzerYaping Liu, R. Alan Harris, R. David Hawkins, R. Scott Hansen, Nezar Abdennur, Mazhar Adli, Mar-

tin Akerman, Luis Barrera, Jessica Antosiewicz-Bourget, Tracy Ballinger, Michael J. Barnes, Daniel Bates, Robert J. A. Bell, David A. Bennett, Katherine Bianco, Christoph Bock, Patrick Boyle, Jan Brinchmann, Pedro Caballero-Campo, Raymond Camahort, Marlene J. Carrasco-Alfonso, Timothy Charnecki, Huaming Chen, Zhao Chen, Jeffrey B. Cheng, Stephanie Cho, Andy Chu, Wen-Yu Chung, Chad Cowan, Qixia Athena Deng, Vikram Deshpande, Morgan Diegel, Bo Ding, Timothy Durham, Lorigail Echipare, Lee Edsall, David Flowers, Olga Genbacev-Krtolica, Casey Gifford, Shawn Gillespie, Erika Giste, Ian A. Glass, Andreas Gnirke, Matthew Gormley, Hongcang Gu, Junchen Gu, David A. Hafler, Matthew J. Hangauer, Manoj Hariharan, Meital Hatan, Eric Haugen, Yupeng He, Shelly Heimfeld, Sarah Herlofsen, Zhonggang Hou, Richard Humbert, Robbyn Issner, Andrew R. Jackson, Haiyang Jia, Peng Jiang, Audra K. Johnson, Theresa Kadlecek, Baljit Kamoh, Mirhan Kapidzic, Jim Kent, Audrey Kim, Markus Kleinewietfeld, Sarit Klugman, Jayanth Krishnan, Samantha Kuan, Tanya Kutyavin, Ah-Young Lee, Kristen Lee, Jian Li, Nan Li, Yan Li, Keith L. Ligon, Shin Lin, Yiing Lin, Jie Liu, Yuxuan Liu, C. John Luckey, Yussanne P. Ma, Cecile Maire, Alexander Marson, John S. Mattick, Michael Mayo, Michael McMaster, Hayden Metsky, Tarjei Mikkelsen, Diane Miller, Mohammad Miri, Eran Mukame, Raman P. Nagarajan, Fidencio Neri, Joseph Nery, Tung Nguyen, Henriette O'Geen, Sameer Paithankar, Thalia Papayannopoulou, Mattia Pelizzola, Patrick Plettner, Nicholas E. Propson, Sriram Raghuraman, Brian J. Raney, Anthony Raubitschek, Alex P. Reynolds, Hunter Richards, Kevin Riehle, Paolo Rinaudo, Joshua F. Robinson, Nicole B. Rockweiler, Evan Rosen, Eric Rynes, Jacqueline Schein, Renee Sears, Terrence Sejnowski, Anthony Shafer, Li Shen, Robert Shoemaker, Mahvash Sigaroudinia, Igor Slukvin, Sandra Stehling-Sun, Ron Stewart, Sai Lakshmi Subramanian, Kran Suknuntha, Scott Swanson, Shulan Tian, Hannah Tilden, Linus Tsai, Mark Urich, Ian Vaughn, Jeff Vierstra, Shinny Vong, Ulrich Wagner, Hao Wang, Tao Wang, Yunfei Wang, Arthur Weiss, Holly Whitton, Andre Wildberg, Heather Witt, Kyoung-Jae Won, Mingchao Xie, Xiaoyun Xing, Iris Xu, Zhenyu Xuan, Zhen Ye, Chia-an Yen, Pengzhi Yu, Xian Zhang, Xiaolan Zhang, Jianxin Zhao, Yan Zhou, Jiang Zhu, Yun Zhu, and Steven Ziegler. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015. ISSN 0028-0836. doi: 10.1038/nature14248. URL http://www.ncbi.nlm.nih.gov/pubmed/25693563.

[47] Wei Xie, Matthew D. Schultz, Ryan Lister, Zhonggang Hou, Nisha Rajagopal, Pradipta Ray, John W. Whitaker, Shulan Tian, R. David Hawkins, Danny Leung, Hongbo Yang, Tao Wang, Ah Young Lee, Scott A. Swanson, Jiuchun Zhang, Yun Zhu, Audrey Kim, Joseph R. Nery, Mark A. Urich, Samantha Kuan, Chia-An Yen, Sarit Klugman, Pengzhi Yu, Kran Suknuntha, Nicholas E. Propson, Huaming Chen, Lee E. Edsall, Ulrich Wagner, Yan Li,

Zhen Ye, Ashwinikumar Kulkarni, Zhenyu Xuan, Wen-Yu Chung, Neil C. Chi, Jessica E. Antosiewicz-Bourget, Igor Slukvin, Ron Stewart, Michael Q. Zhang, Wei Wang, James A. Thomson, Joseph R. Ecker, and Bing Ren. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell*, 153:1134–1148, May 2013.

[48] Danny Leung, Inkyung Jung, Nisha Rajagopal, Anthony Schmitt, Siddarth Selvaraj, Ah Young Lee, Chia-An Yen, Shin Lin, Yiing Lin, Yunjiang Qiu, Wei Xie, Feng Yue, Manoj Hariharan, Pradipta Ray, Samantha Kuan, Lee Edsall, Hongbo Yang, Neil C Chi, Michael Q Zhang, Joseph R Ecker, and Bing Ren. Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature*, 518(7539):350–354, 2015. ISSN 0028-0836. doi: 10.1038/nature14217. URL http://www.nature.com/doifinder/10.1038/nature14217.

[49] Leelavati Narlikar, Noboru J Sakabe, Alexander A Blanski, Fabio E Arimura, John M Westlund, Marcelo A Nobrega, and Ivan Ovcharenko. Genome-wide discovery of human heart enhancers. pages 381–392, 2008. doi: 10.1101/gr.098657.109. URL http://genome.cshlp.org/content/early/2010/01/14/gr.098657.109.short?rss=1.

[50] Kyoung-Jae Won, Iouri Chepelev, Bing Ren, and Wei Wang. Prediction of regulatory elements in mammalian genomes using chromatin signatures. *BMC bioinformatics*, 9:547, 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-547.

[51] Mark D Robinson, Abdullah Kahraman, Charity W Law, Helen Lindsay, Malgorzata Nowicka, Lukas M Weber, and Xiaobei Zhou. Statistical methods for detecting differentially methylated loci and regions. *Frontiers in Genetics*, 5:324, sep 2014. ISSN 1664-8021. doi: 10.3389/fgene.2014.00324. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4165320/.

[52] Hong Ma, Robert Morey, Ryan C. O'Neil, Yupeng He, Brittany Daughtry, Matthew D. Schultz, Manoj Hariharan, Joseph R. Nery, Rosa Castanon, Karen Sabatini, Rathi D. Thiagarajan, Masahito Tachibana, Eunju Kang, Rebecca Tippner-Hedges, Riffat Ahmed, Nuria Marti Gutierrez, Crystal Van Dyken, Alim Polat, Atsushi Sugawara, Michelle Sparman, Sumita Gokhale, Paula Amato, Don P Wolf, Joseph R. Ecker, Louise C. Laurent, and Shoukhrat Mitalipov. Abnormalities in human pluripotent cells due to reprogramming mechanisms. *Nature*, 511(7508):177–83, 2014. ISSN 1476-4687. doi: 10.1038/nature13551. URL http://www.nature.com/doifinder/10.1038/

nature13551{%}5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/25008523.

[53] W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, Haussler, and David. The Human Genome Browser at UCSC. *Genome Research*, 12(6):996–1006, 2002. doi: 10.1101/gr.229102. URL http://genome.cshlp.org/content/12/6/996.abstract.

[54] Matthew D. Schultz, Robert J. Schmitz, and Joseph R. Ecker. 'leveling' the playing field for analyses of single-base resolution dna methylomes. *Trends Genet*, 28:583–585, Dec 2012.

[55] William Perkins, Mark Tygert, and Rachel Ward. An introduction to how chi-square and classical exact tests often wildly misreport significance and how the remedy lies in computers. Jan 2012.

[56] Tim Bancroft, Chuanlong Du, and Dan Nettleton. Estimation of false discovery rate using sequential permutation p-values. *Biometrics*, 69:1–7, Mar 2013.

[57] Heng Li and Richard Durbin. Fast and accurate short read alignment with BurrowsWheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009. doi: 10.1093/bioinformatics/btp324. URL http://bioinformatics.oxfordjournals.org/content/25/14/1754.abstract.

[58] Aaron R Quinlan and Ira M Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010. doi: 10.1093/bioinformatics/btq033. URL http://bioinformatics.oxfordjournals.org/content/26/6/841.abstract.

[59] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, and X Shirley Liu. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):1–9, 2008. ISSN 1474-760X. doi: 10.1186/gb-2008-9-9-r137. URL http://dx.doi.org/10.1186/gb-2008-9-9-r137.

[60] Jennifer Harrow, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L. Aken, Daniel Barrell, Amonida

Zadissa, Stephen Searle, If Barnes, Alexandra Bignell, Veronika Boychenko, Toby Hunt, Mike Kay, Gaurab Mukherjee, Jeena Rajan, Gloria Despacio-Reyes, Gary Saunders, Charles Steward, Rachel Harte, Michael Lin, Cédric Howald, Andrea Tanzer, Thomas Derrien, Jacqueline Chrast, Nathalie Walters, Suganthi Balasubramanian, Baikang Pei, Michael Tress, Jose Manuel Rodriguez, Iakes Ezkurdia, Jeltje van Baren, Michael Brent, David Haussler, Manolis Kellis, Alfonso Valencia, Alexandre Reymond, Mark Gerstein, Roderic Guigó, and Tim J. Hubbard. Gencode: the reference human genome annotation for the encode project. *Genome Res*, 22:1760–1774, Sep 2012.

[61] Chin-Tong Ong and Victor G Corces. CTCF: An Architectural Protein Bridging Genome Topology and Function. *Nature reviews. Genetics*, 15 (4):234–246, apr 2014. ISSN 1471-0056. doi: 10.1038/nrg3663. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4610363/.

[62] Shane Neph, M Scott Kuehn, Alex P Reynolds, Eric Haugen, Robert E Thurman, Audra K Johnson, Eric Rynes, Matthew T Maurano, Jeff Vierstra, Sean Thomas, Richard Sandstrom, Richard Humbert, and John A Stamatoyannopoulos. BEDOPS: high-performance genomic feature operations. *Bioinformatics*, 28(14):1919–1920, 2012. doi: 10.1093/bioinformatics/bts277. URL http://bioinformatics.oxfordjournals.org/content/28/14/1919.abstract.

[63] Yoav Freund and Robert E Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer, 1995.

[64] Jason Ernst and Manolis Kellis. ChromHMM: automating chromatin-state discovery and characterization. *Nat Meth*, 9(3):215–216, mar 2012. ISSN 1548-7091. URL http://dx.doi.org/10.1038/nmeth.1906http://www.nature.com/nmeth/journal/v9/n3/abs/nmeth.1906.html{#}supplementary-information.

[65] Michael M Hoffman, Orion J Buske, Jie Wang, Zhiping Weng, Jeff A Bilmes, and William Stafford Noble. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Meth*, 9(5):473–476, may 2012. ISSN 1548-7091. URL http://dx.doi.org/10.1038/nmeth.1937http://www.nature.com/nmeth/journal/v9/n5/abs/nmeth.1937.html{#}supplementary-information.

[66] L A Pennacchio, N Ahituv, A M Moses, S Prabhakar, M A Nobrega,

M Shoukry, S Minovitsky, I Dubchak, A Holt, and K D Lewis. In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, 444, 2006. doi: 10.1038/nature05295. URL http://dx.doi.org/10.1038/nature05295.

[67] R Kothary, S Clapoff, S Darling, M D Perry, L A Moran, and J Rossant. Inducible expression of an hsp68-lacZ hybrid gene in transgenic mice. *Development (Cambridge, England)*, 105(4):707–714, apr 1989. ISSN 0950-1991 (Print).

[68] Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C. Lin, Peter Laslo, Jason X. Cheng, Cornelis Murre, Harinder Singh, and Christopher K. Glass. 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.

[69] a Liaw and M Wiener. Classification and Regression by randomForest. *R news*, 2(December):18–22, 2002. ISSN 16093631. doi: 10.1177/154405910408300516.

Chapter 4

Dynamic methylome remodeling throughout mammalian fetal development

4.1 Abstract

Genetic studies have revealed an essential role for cytosine DNA methylation in gene regulation. However, its spatiotemporal distribution in the developing embryo remains obscure. Here, we describe analysis of 144 deep-coverage, base-resolution DNA methylomes from profiling of 12 mouse tissues/organs each day from embryonic day 10.5 to birth. We identify 1,808,810 differentially CG methylation regions (CG-DMRs) of which 487,367 show enhancer-like chromatin signatures that are significantly enriched for tissue-related genetic risk factors of human diseases. Strikingly, CG-DMRs predominantly lose CG methylation (mCG) during

fetal development, whereas mCG levels dramatically rise at fetal enhancers after birth. Interestingly, 13%-15% of CG-DMRs are repressed by H3K27me3 during fetal stages, escape postnatal remethylation and store a memory of this fetal regulatory activity in adult tissue epigenomes. Finally, non-CG methylation (mCH) accumulates in the genes bodies of key transcription factors essential for early tissue/organ development, coinciding with their transcriptional repression during late stage fetal development.

4.2 Introduction

Mammalian embryonic development involves spatiotemporal transcriptional regulation, which is mediated by sophisticated orchestration of epigenetic modifications and transcription factor (TF) binding of regulatory DNA elements, primarily enhancers. The accessibility of TFs to regulatory DNA is closely related to both covalent modifications of chromatin and DNA[1, 2, 3, 4, 5].

Cytosine DNA methylation (mC) is an epigenetic modification that plays a critical role in gene regulation[6]. In the mammalian genome, mC occurs predominantly on cytosine followed by guanine (mCG), which is dynamic at regulatory elements in different tissues and cell types[7, 8, 9, 10, 11]. In fact, mCG is able to directly affect the DNA binding affinity of a variety of transcription factors[1, 2, 12, 13, 14] and targeted removal/addition of mCG in promoters is concomitant with increases/decreases in gene transcription[15]. Non-CG methylation (mCH; H = A, C or T) is also present at an appreciable level in embryonic stem cells, oocytes as well as brain, heart, skeletal muscle and a variety of adult tissues[7, 8, 9, 16, 17, 18, 19]. Although its precise function(s) are unknown, mCH abundance directly affects DNA binding of MeCP2, the methyl-binding protein responsible for Rett Syndrome[19, 20, 21, 22].

mC is actively regulated during mammalian development[23, 24]. However, in contrast to pre-implantation embryogenesis[24, 25, 26], data are lacking for the later stages of fetal development, during which anatomical features of the major organ systems are more evident and human birth defects are manifested[27]. To fill this knowledge gap, we used the experimentally tractable mouse embryo as a model system to profile DNA methylation variation. Deep whole-genome bisulfite sequencing[7] was performed to comprehensively profile cytosine DNA methylation in 12 tissue types (in replicate) for 8 development stages starting from embryonic day 10.5 (E10.5) to birth (postnatal day 0, P0; Fig. 1A). The temporal mouse fetal tissue/organ epigenomes (Supplemental Table S1) describe here should inform our understanding the similar cascade of regulatory events occurring during normal human fetal progression as well as developmental disorders. These comprehensive datasets are publically accessible at https://www.encodeproject.org/search/?searchTerm=ecker&type=Experiment&award.rfa=ENCODE3 and
http://neomorph.salk.edu/ENCODE_mouse_fetal_tissues.

4.3 Results

4.3.1 Global and local CG methylation dynamics in fetal tissues/organs.

To assess the cytosine DNA methylation landscape in the developing mouse embryo, 144 high-quality tissue methylomes were produced that cover most of the major organ systems and tissue types derived from the three germ layers. To obtain an overview of these tissue methylomes over development, we first calculated the global mCG level in each tissue/organ at each stage of development (Figure 4.1a-b; Methods). With the notable exception of fetal liver, the genomes of all samples were heavily CG methylated (70% to 82%; liver 60% to 74%). Tissues primarily derived from ectoderm (forebrain, FB; midbrain, MB; hindbrain, HB; neural tube, NT) showed higher mCG levels than other tissue types (heart, HT; craniofacial, CF; limb, LM; kidney, KD; lung, LG; stomach, ST; intestine, IT; liver, LV). Interestingly, large partially methylated domains (PMDs), a genomic feature previously observed in only human cultured cell lines[7, 28], pancreas[8], placenta[29] and cancer samples[30, 31] were found exclusively in mouse fetal liver; PMD formation and dissolution precisely coincided with fetal liver hematopoiesis (Supplemental Note; Figure 4.6).

Despite similar global mCG levels in fetal tissues, massive local mCG differences were observed (Figure 4.1c). We systematically identified 1,808,810 CG differentially methylated regions (CG-DMRs), which are on average 339bp long and cover 22.5% (614 Mb) of the mouse genome (Methods). Comprehensive CG-DMR annotation of all fetal tissues/organs captured 96%(n=272,858) of all previously reported adult mouse tissue CG-DMRs, while adding over 1.5 million new ones (Figure 4.1d). Surprisingly only a minority (8.5% or 153,019) of CG-DMRs overlapped with promoters (+/- 2.5 kb around transcription start sites, TSSs), CpG islands (CGIs) or CGI shores (Figure 4.1e-f; Figure 4.7a-c). The vast majority of CG-DMRs (92.5% or 1,655,791) were located distal to annotated promoters and their underlying DNA sequences showed a high degree of evolutionary conservation (Figure 4.1f-g).

4.3.2 Annotation of methylation variable regulatory DNA

To further classify fetal CG-DMRs, we delineated those genomic regions likely associated with enhancer activity using the REPTILE[32] algorithm which allows enhancer prediction by integration of the fetal tissue mCG data and histone modification data (Gorkin et al companion paper, this issue; Supplemental Table S2). Considering all fetal tissues, except liver, we identified 487,367 CG-DMRs as fetal enhancer-linked CG-DMRs or feDMRs, 85% (415,227) of which are distal to known promoters (Fig. 1f; Methods). feDMRs are evolutionarily conserved and show enhancer-like chromatin signatures including the depletion of mCG and H3K27me3, and the enrichment of H3K4me1 and H3K27ac[7, 33, 34, 35] (Figure 4.1g; Figure 4.7d). Many (106,016) of these enhancers were not previously reported in adult mouse tissues[36] (Figure 4.8a). More than 60% of the DNA elements from VISTA enhancer browser[37] that overlap with feDMRs showed *in vivo* enhancer activity in the predicted tissue or other tissues at E11.5 mouse embryos, and this percentage increased for higher scoring feDMRs (Figure 4.8b). Moreover, tissue-specific feDMRs are enriched for TF binding motifs related to specific tissue function(s) and are near genes in specific tissue-related pathways (Figure 4.8c; Supplemental Table S3).We found that the human orthologs of feDMRs significantly overlapped with the disease/trait-associated single nucleotide polymorphisms identified from genome-wide association studies and showed tissue-specificity (Figure 4.1h; Supplemental Table. S4; Methods). For example, the genetic variants associated with cleft lip are only significantly enriched in craniofacial feDMRs. These tissue-specific enrichments suggest the possibility of generating mouse models of human diseases by introducing the specific disease alleles into feDMRs using genomic editing techniques.

We also identified 221,960 CG-DMRs flanking (within 1kb) the feDMRs but were not predicted as enhancers (Figure 4.1g). We called them flanking distal feDMRs (fd-feDMRs). They are much less conserved than feDMRs and the mCG level of fd-feDMRs is moderately correlated with nearby feDMRs, suggesting that a fraction of them may be the by-products of demethylation in adjacent feDMRs (mean Pearson correlation coefficient = 0.41; Methods). Alternatively, the fd-feDMRs may be bound by pioneer TF(s) which allow opening of chromatin of adjacent feDMRs[38, 39], which is supported by the enrichment of the binding motifs of several known pioneer TFs38 such as FOXA2, GATA3 and PBX1 at fd-DMRs (Supplemental Table S5). Interestingly, the binding motifs of insulator protein CTCF and several transcriptional repressors (e.g. CUX1[40]) are enriched also at fd-feDMRs, indicating a third possibility that fd-feDMRs consist of insulators and silencers (Supplemental Table S5).

Besides the above CG-DMR classes, another type of distal CG-DMR (n = 149,610) is primed distal feDMRs (pd-feDMRs); these display strong CG hypomethylation in at least one tissue sample and are linked to primed fetal enhancers (mCG difference 0.3; Figure 4.9a; Methods). In the tissues where they are hypomethylated, pd-feDMRs showed chromatin signatures resembling primed enhancers[41] (enrichment of H3K4me1 while lacking H3K27ac and H3K27me3; Figure 4.9b). Like feDMRs, pd-feDMRs are also evolutionary conserved (Figure 4.1g). Consistent with their putative role as enhancers, they shared significantly similar TF-binding motif signatures as feDMRs in 9 out of the 12 tissue types (Figure 4.9c; Supplemental Table S6).

The remaining unclassified distal CG-DMRs (868,994) only show subtle CG hypomethylation patterns, indicating that they are likely specific to a small fraction of cells within these complex tissues (mCG difference ¡ 0.3; Methods). Since a functional role cannot yet be assigned, we named this group unexplained CG-DMRs (unxDMR). unxDMRs significantly overlapped with transposable elements (TEs; 58.6%, p-value ¡ 0.001; Methods). Inspired by this observation, we divided unxDMRs into two subgroups: unxDMRs that overlapped with TE (te-unxDMRs) and ones not overlapping (nte-unxDMRs) (Figure 4.1f; Figure 4.9d-e). te-unxDMRs were less evolutionarily conserved compared to flanking regions but they may be a source of novel regulatory elements[42]. Different from te-unxDMRs, genomic sequences underling nte-unxDMRs are as conserved as feDMRs, implying that they may be functional. Indeed, comparing nte-unxDMRs with previously identified mouse regulatory elements, we observed that 10%-45% of the nte-unxDMRs showed open chromatin in purified neurons[43] and/or a variety of mouse cell lines and tissues[44] (p-value ¡ 1e-3; permutation test; Methods). Therefore, nte-unxDMRs are likely regulatory elements active only in rare cell types and their weak hypomethylation profiles are due to the tissue heterogeneity, highlighting the future necessity of cell type-specific or single-cell epigenomic studies.

4.3.3 Distinct temporal mCG dynamics before and after birth

The dominant methylation pattern that emerged during fetal progression was a continuous loss-of-mCG at tissue-specific CG-DMRs, which strongly overlap with predicted enhancers (Figure 4.2a; Figure 4.10a-b; Methods). In striking contrast, the gain-of-mCG methylation at these CG-DMRs mainly occurred after birth (Figure 4.2a; Figure 4.10a). To quantify these changes for each stage interval, we counted loss-of-mCG (mCG decreasing by at least 0.1 in one CG-DMR) and gain-of-mCG events (mCG increasing by at least 0.1 in one CG-DMR) (Figure 4.2b; Methods). During the period from E10.5 to P0, 77% to 95% of the mCG changes involved loss-of-mCG. More than 70% of the loss-of-mCG events occurred between E10.5 to E13.5 in all tissues except in heart (44%; Figure 4.2a; Figure 4.10c). The predominant loss of mCG is likely a shared trend in the majority of cells within these tissues. The mCG level of 44-84% tissue-specific CG-DMRs dropped to below 50% at E14.5, compared to only 31% at E10.5. Since allele-specific methylation is relatively rare[8], the observed methylation dynamics suggest that after E14.5, most of the tissue-specific CG-DMRs are unmethylated in more than half of the cells in a tissue, which may be associated with fluctuations in progenitor cell populations during embryogenesis.

Compared to the loss of mCG, the vast majority (57%-86%) of gain-of-mCG events happened after birth (Figure 4.10d). As a result, 27%-56% of the tissue-specific CG-DMRs become highly methylated in adult tissues (at least 4 weeks old), while the number is 0.3%-15% at birth (P0), reflecting the silencing of fetal regulatory elements (mCG level ¿ 0.6; Figure 4.11a). The lack of gain-of-mCG events during fetal development may be a result of limited de novo CG methylation and/or excessive demethylation activity.

The observed mCG dynamics cannot be explained by the absence of expression cytosines methytransferases *Dnmt1* and *Dnmt3a*, which are highly expressed between E10.5 and E13.5 when major the loss-of-mCG events occur (Figure 4.11b). Furthermore, we also did not find increased expression of *Tet* methylcytosine dioxygenases, involved methylation removal, during the same period. Interestingly, *Tet3* expression is lower in heart, coinciding with its less dynamic mCG during embryogenesis. Absence of gain-of-methylation events until the postnatal period may involve translational or posttranslational regulation of these enzymes or the absence of cofactors or proteins that target them or possibly other unknown control mechanisms.

4.3.4 Dynamic CG methylation is associated with the remodeling of enhancer-related chromatin states

To further pinpoint the timing of CG-DMR remethylation and its relationship with enhancer activity, we included methylation data from adult frontal cortex[9] as well as H3K27ac data adult forebrain[45] (Supplemental Table S2). We then clustered forebrain-specific CG-DMRs into 8 groups based on the mCG and H3K27ac dynamics across both fetal and adult stages (Figure 4.2e; Figure 4.11c; Methods). CG-DMRs in each group were located nearby genes related to distinct brain related processes (Figure 4.11d). Despite the fact that frontal cortex is only part of forebrain, the CG-DMRs hypomethylated in P0 forebrain were also hypomethylated in frontal cortex at the first postnatal week (P1wk; Figure 4.2E). Between postnatal 1 and 2 weeks, methylation of forebrain-specific CG-DMRs increased dramatically and become even further methylated as the tissue matures (Figure 4.10e). Using these additional datasets, we were able to refine the timing of remethylation in the frontal cortex.

We then asked how the mCG dynamics at tissue-specific CG-DMRs were associated with their enhancer activity (approximated by H3K27ac abundance) during development. Interestingly, although the temporal depletion of mCG was not necessarily related to high H3K27ac enrichment (e.g. C3, C5 and C6), high methylation level is indicative of low H3K27ac (Figure 4.2e-f). Across all tissues except liver, the tissue-specific CG-DMRs that are highly CG methylated (mCG level ¿ 0.6) are less frequent to show strong enrichment of H3K27ac compared to ones that are lowly or moderately CG methylated (Figure 4.2f). This trend is independent of development stages. Collectively, these results suggest that depletion of mCG in regulatory elements associated with enhancer activity and the decreasing methylation level at CG-DMRs during development may prime flexible regulation of enhancer activity.

4.3.5 Correlation between methylation dynamics and gene co-expression networks

We investigated the association between differential mCG and the transcription of genes in different pathways, using the RNA-seq data from matched samples

(Wold et al. see companion paper, this issue). By applying an unsupervised, weighted correlation network analysis (WGCNA)[46] method, we identified 33 co-expressed gene clusters (co-expression modules, CEMs) and we calculated eigengenes to summarize the expression profile of genes within modules (Figure 4.3a-b; Figure 4.12a-b; Methods). Genes sharing similar expression profiles are likely regulated by a common mechanism and/or involved in the same pathway. For example, CEM12 contains genes that are highly expressed in early developmental stages but are down regulated as tissues mature (Figure 4.3c; Figure 4.12b). Genes in CEM12 are significantly related to cell cycle, matching our knowledge that cells become post-mitotic in mature tissues. Similarly, genes in CEM3 are related to chromatin modification and are lowly expressed in heart, which showed less mCG dynamics relative to other tissues (Figure 4.3b-c; Figure 4.12c). Overall, genes in CEMs are associated with different pathways and/or biological processes (Figure 4.12d; Supplemental Table S7).

To understand how mCG profiles and enhancer activity of regulatory elements (feDMRs) are associated with the expression of genes in CEMs, we first inferred feDMRs target genes based on genomic distance and linked each feDMR to its neighboring gene (Methods). Next, for each CEM, we correlated its eigengene expression with the average mCG and enhancer activity of the feDMRs that were linked to the genes in that CEM (Methods). We used enhancer scores to approximate enhancer acitivity because higher scoring feDMRs are more likely to display *in vivo* enhancer activity (Figure 4.8b).

To tease out the effect of tissue type and development, we calculated the correlations separately for tissue-specific expression and temporal expression (Methods). Given a developmental stage, a higher level of mCG at feDMRs was negatively correlated with the tissue-specific eigengene expression, consistent with its known repressive role (Figure 4.3d-e). In contrast, the enhancer score of feDMRs was positively correlated with tissue-specific transcription, implying that enhancers are likely the drive of tissue-specific expression and this score provides a good approximation of enhancer activity (Figure 4.3e-e). For example, genes related to synaptic transmission (CEM32) are highly expressed in neuronal tissues, while their neighboring feDMRs are depleted of mCG and show high enhancer scores (Figure 4.3d; Figure 4.12d). Such trends hold when we calculated the correlations for all modules (Figure 4.3e; Methods).

Next, for a given tissue type, we calculated the correlation across development. Because CG methylation generally decreased at feDMRs in fetal tissues over, mCG only showed marginally better anticorrelation with temporal transcription than that by chance (Figure 4.2a; Figure 4.3f-g). In contrast, the enhancer score remains positively correlated with the temporal expression, implying that enhancer activity is the driver of temporal gene expression (Figure 4.3f-g).

4.3.6 Epigenetic memory of the fetal enhancers in adult tissues

Although the majority of CG-DMRs gain methylation after birth, not all of them become hypermethylated in adult tissues (Figure 4.11a). The regions hypomethylated in adult tissues have been termed either adult vestigial enhancers (AD-V enhancers) or adult active enhancer (AD-A enhancers)[11]: Compared to AD-A enhancers, AD-V enhancers are depleted of enhancer-like chromatin marks and enriched for H3K27me3. However, both types are CG hypomethylated, evolutionarily conserved and enriched for TF binding motifs[11]. Also, AD-V enhancers were able to drive gene transcription *in vivo* in fetal tissues[11], indicating that they are likely a remnant of fetal enhancers in adult tissues. However, this hypothesis has never been systematically investigated.

Taking advantage of the temporal dimension of our dataset, we traced the epigenetic signatures of the AD-V and AD-A enhancers identified in adult heart, intestine and kidney from a previous study[11] (Methods). Although examples of AD-V and AD-A enhancers exist that are enriched for enhancer-like chromatin signatures and show enhancer activity at fetal stages, only 4%-14% of AD-V enhancers overlapped with feDMRs compared to 70%-84% for AD-A enhancers (Figure 4.4a-b; Figure 4.13a). To further interrogate this, we narrowed our scope to CG-DMRs that overlapped AD-V (AD-V CG-DMRs) or AD-A enhancers (AD-A CG-DMRs) and found that, consistently, 19%-32% of the AD-V CG-DMRs are enriched for H3K27ac and predicted as enhancers (AD-V feDMRs), fewer than AD-A CG-DMRs (68%-81%; Figure 4.13b). Such differences may be a result of the transient fetal activity of AD-V enhancers, which is not evident in the sampled stages. Indeed, the enhancer score of the AD-V feDMRs is more dynamic than AD-A feDMRs (Figure 4.13c; p-value ¡ 0.01, Mann-Whitney test). Furthermore, AD-V feDMR target genes expressed more dynamically (Figure 4.13d; p-value ¡ 0.01, Mann-Whitney test). In addition, AD-V feDMRs showed lower enhancer scores and the expression levels of their target genes was also lower than AD-A feDMRs (Figure 4.13e-f). These observations indicate that the fetal activity of AD-V enhancers is more dynamic and is more difficult to detect.

To understand the regulation of AD-V enhancers, we studied the epigenetic marks likely responsible for AD-V non-feDMR repression throughout fetal development. On average, only 44% of the AD-V non-feDMRs were heavily CG methylated (mCG level ¿ 0.6) on fetal stages, lower than AD-A non-feDMRs (68%; Figure 4.4c-d; Figure 4.14a-c). In contrast, AD-V non-feDMRs showed stronger H3K27me3 enrichment compared to AD-A non-feDMRs (Figure 4.14d). As H3K27me3 and mCG are complementary gene silencing mechanisms[47], the lowly methylated AD-V non-feDMRs may be repressed by H3K27me3. Indeed, lowly methylated AD-V non-feDMRs (mCG level ¡ 0.6) showed stronger H3K27me3 enrichment than highly methylated ones (mCG level ¿= 0.6; Figure 4.14e). Enrichment of H3K27me3 at AD-V non-feDMRs may result in the formation of vestigial enhancer signatures by preventing CG remethylation. Expanding the analysis on all fetal tissue-specific CG-DMRs, we found that 13%-15% (10,757-13,948) of them are repressed by H3K27me3 at fetal stages and show vestigial epigenomic state in adult tissue (Methods). Our results support a model in which H3K27me3 and mCG may each repress distinct subsets of AD-V non-feDMRs, whereas mCG is the primary mechanism for silencing AD-A non-feDMRs, which undergo dramatic demethylation after birth and become active in the adult.

4.3.7 Intragenic non-CG methylation is associated with gene repression

Non-CG methylation, a less well understood form of cytosine DNA methylation in mammals, is present in most adult tissues and is the dominant form of cytosine methylation in human neurons[17]. No mechanism is known to actively remove mCH, although passive dilution can occur during cell replication[17]. Surprisingly, we found that mCH accumulates to detectable levels in nearly all fetal tissues during their developmental trajectories (Figure 4.5a). Interestingly, the timing of mCH accumulation varies in different tissues (Figure 4.5a). Heart tissue showed traceable mCH at as early as E10.5, when mCH was not observable in other tissues. Three brain tissues, forebrain, midbrain and hindbrain, also showed different rates of mCH accumulation. Previous studies revealed that mCH is preferentially deposited at 5-CAG-3 context in embryonic stem cells and 5-CAC-3 context in brain, heart, skeletal muscle and a variety of adult tissues[7, 8, 9, 16, 17, 18]. In all fetal tissues mCH was found in the CAC context and the significance of this 3 base specificity increased as tissues mature, implying a similar pathway (Dnmt3A) as is responsible for mCH in adult tissues[17] (Figure 4.15a).

mCH was not uniformly distributed across the genome but preferentially accumulated in genomic loci that we termed as mCH domains, genomic regions that showed higher mCH level than flanking sequences (Figure 4.5b). We systematically identified 384 mCH domains that averaged 255kb in length, encompassing 98Mb in total (Methods). Strikingly, 92% (355 out of 384) of the mCH domains and 61% of bases overlapped with annotated gene bodies (p-value ¡ 0.001; Methods). A highly significant fraction (22%) of these mCH domain genes (e.g. *Pax3*) encode transcription factors (Figure 4.5b; 128 out of the 581 genes, p-value ¡ 0.001; Methods).

To explore the tissue and temporal specificity of mCH accumulation, we used k-means clustering to group mCH domains into 5 clusters based on methylation dynamics (Figure 4.5b-c; Figure 4.15b; Methods). mCH domains in clusters 1 and 3 acquire mCH in all tissues and are enriched for genes related to embryo development (Figure 4.5c-e; Figure 4.15b; Supplemental Table S8). mCH domains in cluster 4, also accumulate in all tissues but are more dramatic than clusters 1 and 3, and are significantly enriched for genes with neuron differentiation functions. In contrast to these ubiquitous mCH domains, cluster 2 gains mCH most evidently in heart where as those in cluster 5 show brain-specific mCH accumulation, overlapping genes that are enriched for functions related to axon guidance (Supplemental Table S8). Since distinct mCH landscapes have been found in different cell types in brain[9, 43], the observed fetal tissue-specific mCH dynamics may indicate that in different tissues, mCH likely accumulates in distinct (combinations of) cell types.

We then asked how the mCH accumulation is associated with gene expression as mCAC in the gene body was found anticorrelated with gene expression[17]. Indeed, as methylation accumulates in mCH domains, the genes within tend to be repressed compared to genes outside mCH domains in late developmental stages, especially at P0 (Figure 4.5f; Figure 4.15c). Since mCH domains are enriched for TFs and other genes related to tissue/organ or embryo development, our data suggests that mCH may be associated with silencing pathways of early development.

Interestingly, mCH domains are enriched for feDMRs compared to flanking regions in a tissue-specific manner (Figure 4.5g). mCH domains that showed ubiquitous mCH accumulation (C1 and C3) are enriched for feDMRs of all tissue types. For mCH domains in tissue-specific clusters, the feDMR enrichment is found only for heart (C2) and brain-related tissues (C5) (Figure 4.5g). For cluster C4, feDMRs enrichment was observed in all tissues but it is most evident for feDMRs in brain-related tissues. feDMRs in mCH domains exhibit a decreasing trend of enhancer activity as development proceeds (Figure 4.15d). Moreover, these feDMRs tend to become hypermethylated in late stages compare with feDMRs that lie outside mCH domains (Figure 4.15e). These observations indicate that mCH accumulation predicts the future silencing of regulatory elements, consistent with recently reported findings for human cerebral organoids[48]. Collectively, we found mutual associations between several mCH domains features including: mCH accumulation, enrichment for genes related to the tissues that acquire mCH, enrichment of feDMRs, down-regulation of gene transcription, as well as decreasing enhancer activity. Delineating the mechanism(s) that drive these associations will provide new insights into mCH regulation and the potential involvement of non-CG methylation in transcriptional regulation.

4.4 Conclusion

In this study we describe the generation and analysis of a comprehensive collection of base-resolution, genome-wide maps of cytosine methylation for 12 fetal tissue types from 8 developmental stages of mouse embryogenesis. By integrating DNA methylation, histone modification and RNA sequencing data from the same tissue samples, we annotated millions of methylation variable elements including prediction of fetal enhancers and sets of transcription factors that bind them. Because of the temporal nature of these data, we uncovered surprisingly simple mCG dynamics at predicted DNA regulatory regions where during early stages of fetal development methylation decreases in all tissues until birth at which time methylation at predicted fetal regulatory elements dramatically rises. In spite of the tissue heterogeneity, such dynamics suggest a plausible regulatory principal whereby stable repressive mCG is removed to enable a mode of more flexible transcriptional regulation (e.g. histone modifications). In addition, we reveal that the formation of previously identified adult vestigial enhancers[11] is likely due to the antagonistic interaction between H3K27me3 and mCG, as an example of crosstalk between epigenetic modifications during development. Also, our findings extend current knowledge of methylation in a non-CG context, an understudied type of DNA methylation. We observed that during fetal development there is preferential accumulation of mCH at genomic locations each hundreds of kilobases in size, a novel genomic feature we termed mCH domains. Genes that lie in mCH domains tend to become down regulated as mCH accumulates in the later stages of fetal development . Though its function remains debatable, *in vivo* and *in vitro* studies indicated that mCH directly increases the binding affinity of MeCP2[20, 21, 22], mutation of which leads to Rett Syndrome. Gene-rich mCH domains are likely enriched for MeCP2 binding which may be involved in the observed transcriptional repression. Our study highlights the power of temporal tissue epigenome maps to uncover regulatory element dynamics in fetal tissues during in utero development. These datasets provide a valuable resource for studies of fundamental questions about gene regulation during mammalian tissue/organ development as well as knowledge about the possible origins of human developmental diseases.

4.5 Acknowledgements

We thank Drs. Shao-shan Carol Huang, Liang Song and Eran Mukamel for critical comments. Y.H. is supported by the H.A. and Mary K. Chapman Charitable Trust. D.U.G is supported by the A.P. Giannini Foundation and NIH IRACDA K12 GM068524. J.R.E is an Investigator of the Howard Hughes Medical Institute and is supported by grants from the Gordon and Betty Moore Foundation (GBMF3034), the National Institutes of Health (R01 MH094670 and U01 MH105985), and the California Institute for Regenerative Medicine (GC1R-06673-B). This work was supported by the National Institutes of Health ENCODE Project (U54 HG006997).

Chapter 4, in full, is a reprint of a manuscript to be submitted to Nature. Yupeng He, Manoj Hariharan, David U. Gorkin, Diane E. Dickel, Chongyuan Luo, Rosa G. Castanon, Joseph R. Nery, Ah Young Lee, Brian A. Williams, Diane Trout, Henry Amrhein, Rongxin Fang, Huaming Chen, Bin Li, Axel Visel, Len A. Pennacchio, Bing Ren and Joseph R. Ecker. Dynamic methylome remodeling throughout mammalian fetal development. In Preparation. The dissertation author was primary investigator and author of this paper.

4.6 Supplemental Note - mCG landscape remodeling in fetal liver

Distinct from the hypermethylated genome of all other tissues, the liver genome underwent drastic global demethylation from E11.5 to E14.5, and remained hypomethylated till E16.5, after which it returned to hypermethylated state at P0 (Figure 4.1B). The hypomethylated liver genome, present during E12.5 to E16.5, displayed a partially methylated domains (PMDs) signature, a methylation feature previously observed in human cultured cell lines[7, 28], pancreaspancreas[8], placenta[29] and cancer samples[30, 31]. PMDs are large genomic regions (typically greater than 100kb) that are lowly CG methylated (Figure 4.6A). We systematically identified PMDs in liver samples from all stages (Methods). Strikingly, from E14.5 to E16.5, PMDs covered more than half of the genome and the coverage shrunk dramatically afterwards (Figure 4.6B). We found that PMDs identified in E15.5 displayed hypomethylation in all liver samples and covered almost all PMDs from other stages (Figure 4.6C-D). These results indicate that the PMDs identified at different fetal stages are essentially identical and the different PMDs calls were due to various signal-to-noise ratios. Therefore, we defined the PMDs identified in E15.5 liver as liver PMDs (n = 4,578; average size = 338kb).

Mouse liver PMDs share all molecular signatures of the PMDs identified in human fibroblast cell lines, normal and cancer tissues[7, 8, 30, 31]: First, mouse PMDs are enriched for H3K9me3 and H3K27me3 and are depleted of H3K27ac (Figure 4.6E). Second, mouse PMDs tend to be replicated during the later stages of the cell cycle and strongly overlap with lamina-associated domains (Figure 4.6F;

p-value ¡ 0.001; permutation test; Methods). Furthermore, we found that genes overlapping with mouse PMDs tend to have lower expression compared to genes outside PMDs, which was also reported by Schultz et al for human pancreas[8]. These shared properties indicate that human and mouse PMDs are likely identical genome feature and their presence may be due to similar mechanism, likely the failure of mCG maintenance in rapidly dividing cells[30].

The presence of PMDs in fetal liver coincides with hematopoiesis[49, 50]. Hematopoiesis initiates at E11.5, while liver genome remains hypermethylated (Figure 4.1B). Then, hematopoietic expansion occurs between E12.5 and E14.5, during which the liver genome underwent demethylation and PMDs became evident (Figure 4.1B; Figure 4.6B and D). The increasing number of rapidly dividing cells during this expansion period may explain the formation of PMDs. After E15.5, the hematopoiesis starts to disappear although the liver tissue genome is not fully remethylated until P0 (Figure 4.6B).

4.7 Methods

4.7.1 Data Availability

All whole-genome bisulfite sequencing (WGBS) data of embryonic tissues are available in ENCODE portal (https://www.encodeproject.org/) and most of them can be accessed from GEO (Supplemental Table 1). All other data used in this study, including chromatin immunoprecipitation with massively parallel DNA sequencing (ChIP-seq), RNA-seq and additional WGBS data, are available in ENCODE portal and/or GEO (Supplemental Table 2).

4.7.2 Abbreviations

- AD: adult
- CEM: co-expression module
- mC: cytosine DNA methylation
- mCG: CG methylation
- mCH: non-CG methylation
- TF: transcription factor
- H3K4me1: Histone 3 lysine 4 monomethylation
- H3K4me3: Histone 3 lysine 4 trimethylation)
- H3K27me3: Histone 3 lysine 27 trimethylation
- H3K27ac: Histone 3 lysine 27 acetylation
- WGBS: whole-genome bisulfite sequencing
- REPTILE: Regulatory Element Prediction based on TIssue-specific Local Epigenetic marks
- GWAS: genome-wide association study
- SNP: single nucleotide polymorphism
- TPM: transcripts per million
- WGCNA: weighted gene co-expression network analysis

- RPKM: Reads Per Kilobase per Million mapped reads

Genomic features

- CG-DMR: differentially CG methylation region
- feDMR: fetal enhancer linked CG-DMR
- fd-feDMR: flanking distal feDMR
- pd-feDMR: poised distal feDMR
- unxDMR: unexplained CG-DMR
- te-unxDMR: transposable element overlapping unxDMR
- nte-unxDMR: non transposable element overlapping unxDMR
- TSS: transcription start sites
- CGI: CpG island
- PMD: partially methylated domain
- AD-A enhancer: adult active enhancer
- AD-V enhancer: adult vestigial enhancer

Tissues/organs

- FB: forebrain
- MB: midbrain

- HB: hindbrain
- NT: neural tube
- HT: heart
- CF: craniofacial
- LM: limb
- KD: kidney
- LG: lung
- ST: stomach
- IT: intestine
- LV: liver

4.7.3 Tissue Collection And Fixation

See Supplemental File 1-2 for details.

4.7.4 MethylC-seq Library Construction

MethyC-seq library was constructed as described previously[8] and the detailed protocol can be found in Urich et al[51]. Illumina HiSeq 2500 was used to sequence the libraries and generate 100 or 130 bases single-ended reads.

4.7.5 Mouse Reference Genome Construction

For all analyses in this study, we used mm10 as reference genome, which includes 19 autosomes and two sex chromosomes (corresponding to the mm10-minimal reference in ENCODE portal, https://www.encodeproject.org/). The fasta files of mm10 were downloaded from UCSC genome browser (Jun 9 2013)[52].

4.7.6 WGBS Data Processing

All WGBS data were processed on mm10 mouse reference genome exactly as previously described[53]. The reference for WGBS processing also included lambda genome as control to estimate sodium bisulfite non-conversion rate. The pipeline, methylpy, is available on github (https://github.com/yupenghe/methylpy). Briefly, cytosines on WGBS reads were first converted to thymines. The converted reads were then aligned by bowtie (1.0.0) onto the forward strand of C-T converted reference genome and the reversed strand of G-A converted reference genome, separately. We filtered out reads that were not uniquely mapped or were mapped to both converted genomes. Next, PCR duplicates were also removed. Last, methylpy counted the methylated basecalls (cytosines) and unmethylated basecalls (thymines) at each cytosine.

4.7.7 Calculation Of Methylation Level

Methylation level was computed to measure the intensity and degree of DNA methylation of single cytosine or genomic region. Methylation level is defined as the ratio of the sum of methylated basecall counts over the sum of both methylated and unmethylated basecall counts at one cytosine or across sites in a given region[54] subtracting sodium bisulfite non-conversion rate. The sodium bisulfite non-conversion rate is defined as the methylation level of the lambda genome. We calculated this metric for cytosines on CG context and CH context (H=A, C or T). The former is called CG methylation (mCG) level or mCG level. Similarly, the latter is called CH methylation (mCH) level or mCH level.

4.7.8 ChIP-seq Data Processing

ChIP-seq data were processed using the ENCODE uniform processing pipeline for ChIP-seq: briefly, reads were first mapped to the mm10 reference using bwa[55] (version 0.7.10) with parameters -q 5 -l 32 -k 2. Next, Picard tool (http://broadinstitute.github.io/picard/, version 1.92) removed PCR duplicates with parameters REMOVE_DUPLICATES=true. For each histone modification mark, we represented it as continuous enrichment values of 100bp bins across the genome. The enrichment was defined as the RPKM (Reads Per Kilobase per Million mapped reads) of ChIP subtracting input. The enrichment across the genome was calculated using bamCompare in Deeptools2[56] with options –binSize 100 –normalizeUsingRPKM –extendReads 300 –ratio subtract. For the ChIP-seq data of EP300, we used MACS[57] (1.4.2) to call peaks given default parameters.

4.7.9 RNA-seq Data

Processed RNA-seq data of all fetal tissues from all stages was downloaded from ENCODE portal (https://www.encodeproject.org/; Supplemental Table S2). To further validate the finding regarding transcriptome, we generated additional

RNA-seq data for forebrain, midbrain, hindbrain and liver. We first extracted total RNA using RNeasy Lipid tissue mini kit from Qiagen (cat no.#74804). Then, we used Truseq Stranded mRNA LT kit (Illumina, RS-122-2101 and RS-122-2102) to constructed stranded RNA-seq libraries on 4ug of the extracted total RNA. Illumina HiSeq 2500 were used to sequence the libraries and generate 100 bases single-ended reads.

4.7.10 RNA-seq Data Processing And Gene Expression Quantification

The RNA-seq data was processed using ENCODE RNA-seq uniform processing pipeline. Briefly, RNA-seq reads were mapped to mm10 mouse reference using STAR[58] aligner (version 2.4.0k) with Gencode M4 annotation[59]. We quantified the gene expression levels using RSEM (version 1.2.23)[60], expressed as transcripts per million (TPM). For all downstream analysis, we filtered out non-expressed genes and only retain only the genes that showed non-zero TPM in at least 10% of samples.

4.7.11 Genomic Features of Mouse Reference Genome

We used GENCODE M4 gene annotation[59] in this study, which is used by ENCODE for mouse data. CG island (CGI) annotation was downloaded from UCSC genome browser (Sep 5, 2016)[52]. CGI shores are defined as the upstream 2kb and downstream 2kb regions along CGIs. Promoters are defined as regions from - 2.5kb to +2.5kb around transcription start sites (TSSs). CGI promoters are one with any overlap with CGIs and the remaining promoters are non-CGI promoters. We also obtained a list of mappable transposable elements (TEs) using the below procedure. RepeatMasker annotation of mm10 mouse genome was downloaded from UCSC genome browser (Sep 12, 2016)[52]. The annotation includes 5,138,231 repeats. We acquired the transposon annotation by selecting only the repeats belong to one the below repeat classes (repClass): DNA, SINE, LTR or LINE. Then, we excluded any repeat elements with question mark in their name (repName), class (repClass) or family (repFamily). In the remaining 3,643,962 transposons, we further filtered out elements that contained less than 2 CG sites or less than 60

4.7.12 CG Differentially Methylated Region (CG-DMRs)

We identified CG-DMRs using methylpy (https://github.com/yupenghe/methylpy) as previously described[53]. Briefly, we first called CG differentially methylated sites (CG-DMSs) and then merged them into blocks if they both show similar sample-specific methylation patterns and are within 250bp. Last, we filtered out the blocks that contain less than three CG-DMSs. In this procedure, we combined the data of the biological replicates for each tissue and we only considered the data of non-liver tissues due to the global hypomethylation in liver genome. We overlapped the CG-DMRs with CG-DMRs identified in Hon et al[11] using intersectBed from bedtools[61]. The mm9 coordinates of the CG-DMRs from Hon et al. were first, mapped to mm10 using liftOver3 with default parameters. An overlap in one CG-DMR list is defined as a CG-DMR with at least one base with any CG-DMRs in the other list. The result is shown in Figure 4.1d.

4.7.13 Identification of tissue-specific CG-DMRs

For each tissue type, we defined tissue-specific CG-DMRs as the CG-DMRs that showed hypomethylation in the tissue samples from any stages between E10.5 to P0. Hypomethylation is only meaningful with a baseline. Inspired by how baseline was defined in an outlier detection algorithm[62], we defined baseline mCG level of each CG-DMR across tissue samples as the mean of the bulk, which is defined as the values in the narrowest mCG level interval that includes at least half of the samples. Specifically, x_s^i is the mCG level of CG-DMR $i(i = 1, , M)$ in tissue sample $s(s = 1, , N)$. Assuming the samples are ordered such that $x_1^i \leq x_2^i \leq x_s^i \leq x_N^i$, the baseline is defined as the $b_i = \sum_{s=a}^{a+\lceil N/2 \rceil} x_s^i$, where a is the sample index such that $x_{a+\lceil N/2 \rceil}^i - x_a^i$ is minimized, i.e. $a = argmin_t(x_{t+\lceil N/2 \rceil}^i - x_t^i)$. $\lceil N/2 \rceil$ is defined as the smallest integer that is greater than $N/2$. Last, we defined hypomethylated samples as the samples in which the mCG level at CG-DMR i is at least 0.3 smaller than baseline b_i, i.e. $s|(x_s^i - b_i) \leq -0.3$. Then, CG-DMR i is specific to these tissues. Liver data was not included in this analysis and we excluded any CG-DMRs that had zero coverage in any of the non-liver samples. In total, 402 (0.02%) CG-DMRs were filtered out.

4.7.14 Linking CG-DMRs To Genes

We linked CG-DMRs to their target genes based on genomic distance. First, we only considered expressed genes, which showed non-zero TPM in at least 10% of all fetal tissue samples. Next, we obtained the TSSs of the expressed genes and paired each CG-DMR with the closest TSS using closestBed from bedtools[61]. In this way, we inferred the target gene of each CG-DMR and this map was used in all the analyses in this study.

4.7.15 Predicting Fetal Enhancer-linked CG-DMRs

REPTILE[32] algorithm was used to identify the CG-DMRs that showed enhancer-like chromatin signature and were likely to act like enhancers. We called them fetal enhancer-like CG-DMRs or feDMRs. REPTILE uses random forest classifiers to learn and then distinguish the epigenomic signatures of enhancers and genomic background. One unique feature of REPTILE is that by incorporating the data of additional samples (as outgroup/reference), REPTILE is able to employ epigenomic variation information to improve enhancer prediction. In this analysis, we ran REPTILE using the data of CG methylation (mCG) and six histone marks (H3K4me1, H3K4me2, H3K4me3, H3K27ac, H3K27me3 and H3K9ac). REPTILE enhancer model was trained as previously described[32]. Briefly, CG-DMRs were called across the methylomes of mouse embryonic stem cells (mESCs) and all eight E11.5 mouse tissues. CG-DMRs were required to contain at least 2 CG-DMSs and they were extended 150bp from each direction. The REPTILE model was trained on the mESC data using E11.5 mouse tissues as outgroup. Data of mCG and six histone modifications are available for these samples. The training dataset consists of 5000 positive instances (putative known enhancers) binding and 35,000 negative instances. Positives were 2kb regions centered at the summits of top 5,000 EP300 peaks in mESCs. Negatives include randomly chosen 5,000 promoters and 30,000 2kb genomic bins. The bins have no overlapped with any positives or promoters. REPTILE learned the chromatin signatures of these positive and negative instances, based on which it identified the CG-DMRs that are likely to be enhancers.

Next, using this model, we applied the REPTILE to delineate feDMRs from the 1,808,810 CG-DMRs identified across all non-liver tissues. The feDMRs were predicted for each sample based on the data of mCG and six histone marks, while the remaining non-liver samples were used as outgroup. In REPTILE, the random forest classifier for CG-DMR assigns a confidence score ranging from 0.0 to 1.0 to each CG-DMR in each sample. The score corresponds to the fraction of decision trees in the random forest model that vote in favor of the CG-DMR to be an enhancer. Previous benchmark showed that higher the score, more likely the CG-DMR shows enhancer activity[32]. We named the confidence score as putative enhancer activity (PEA). In each tissue sample, feDMRs are the CG-DMRs with PEA greater than 0.3. feDMRs were also defined for each tissue type as the CG-DMRs that were identified as feDMR in at least one tissue sample of that tissue type. For example, if a CG-DMR was predicted as feDMR only in E14.5 forebrain, it is a forebrain-specific feDMR. We overlapped the feDMRs with putative enhancers from Yue et al[36]. We downloaded the center of putative enhancers in each of the tissues and cell types from http://yuelab.org/mouseENCODE/predicted_enhancer_mouse.tar.gz. Then, we defined putative enhancers as +/- 1kb regions around the centers. Putative enhancers from different tissues and cells types were combined and merged if they were overlapped. The merged putative enhancers (mm9) were mapped to mm10 reference using liftOver[52]. Finally, intersectBed from bedtools[61] were used to overlap feDMRs with these putative enhancers.

4.7.16 Enhancer score and in vivo enhancer activity

To estimate the likelihood that a feDMR with certain enhancer score actually displays enhancer activity in vivo, we downloaded enhancer validation data from VISTA enhancer browser[37] and validated feDMRs with DNA elements (VISTA elements) that were experimentally tested for enhancer activity in E11.5 embryo. We used this dataset to validate feDMRs predicted in six E11.5 tissues (forebrain, midbrain, hindbrain, heart, limb and neural tube), where at least 30 validated VISTA elements (enhancers) are available. To do this, in each E11.5 tissue, we first overlapped feDMRs predicted in that tissue with VISTA elements and picked out VISTA elements that fully contained at least one feDMR. Then, we calculated the fraction of feDMR overlapping VISTA elements that displayed enhancer activity in the predicted tissue, any tissue(s) or nowhere. The results are shown in Figure 4.7A.

4.7.17 Enriched transcription factor (TF) binding motifs in tissue-specific feDMRs

To identify TF motifs enriched in feDMRs, we scanned the genome to delineate TF motif occurrences as previously described[43]. Briefly, we downloaded TF binding position weight matrices (PWMs) from the MEME motif database (v11, 2014 Jan 23. motif sets chen2008, hallikas2006, homeodomain, JASPAR_CORE_2014_vertebrates, jolma2010, jolma2013, macisaac_theme.v1, uniprobe_mouse, wei2010_mouse_mw, wei2010_mouse_pbm, zhao2011). Then, FIMO[63] was used to scanned the genome to identify TF motif occurrences using options –output-pthresh 1E-5 –max-stored-scores 500000. Next, we performed hypergeometric test to identify significant motif enrichments. For each tissue type, we calculated the motif enrichment in the feDMRs of that tissue (foreground) against the feDMRs of other tissues that were not overlapped with foreground. In this analysis, we extended the average size of both foreground and background to 400bp to avoid bias due to size difference. For a given tissue t, the total number of foreground and background feDMRs is $N_{f,t}$ and $N_{b,t}$, respectively, and $N_t = N_{f,t} + N_{b,t}$ is the total number of feDMRs. For a given TF binding motif m, TF motif occurrences are overlapped with $n_{f,t,m}$ foreground and $n_{b,t,m}$ background feDMRs, while $n_{t,m} = n_{f,t,m} + n_{b,t,m}$ is the total number of overlapping feDMRs. The probability of observing $n_{f,t,m}$ or more overlapping foreground feDMRs (p-value) is defined as:

$$P(Xn_{f,t,m}|N_{f,t}, n_{f,t,m}, N_{b,t}, n_{b,t,m}) = \sum_{x=n_{f,t,m}}^{n_{t,m}} \frac{\binom{N_{f,t}}{x}}{\binom{N_{b,t}}{n_{t,m}-x}\binom{N_t}{n_{t,m}}} \tag{4.1}$$

For each tissue type, we performed this test for all motifs (n=532). Then, the p-values of each tissue were adjusted using Benjamini-Hochberg method and the motifs were called as significant if they passed 1% FDR cutoff. Last, we excluded any TF-binding motifs whose TF expression level was less than 10 TPM.

4.7.18 Enrichment of GWAS SNPs in feDMRs

Genome-wide association study (GWAS) SNPs were first downloaded from GWAS Catalog[64] (gwas_catalog_v1.0-associations_e86_r2016-11-28.tsv). We then filtered out the SNPs with missing coordinate or missing p-value information as well as ones whose p-value is greater than 5x10-8, ending up with 13,470 SNPs. Next, we used liftOver from UCSC genome browser[52] to convert their coordinates from hg38 human reference genome to hg19, after which 4 SNPs were excluded. We further selected the SNPs that could be lifted over to mm10 mouse reference genome. In the end, 7,052 GWAS SNPs that are conserved between human and mouse were included in following analysis. Next, to obtain the human orthologs of CG-DMRs, we used liftOver to map CG-DMRs (mm10) to hg19, requiring that at least 50% of the bases in CG-DMR can be mapped to hg19 (using option -minMatch=0.5). In total, 1,034,801 out of 1880810 DMRs (55%) left. Finally, we overlapped the human orthologs of feDMRs of each tissue/organ with the GWAS SNPs and tested for enrichment using one-tailed hypergeometric test. Specifically, for each tissue/organ, we overlapped the GWAS SNPs with the human orthologs of distal feDMRs in that tissue/organ (foreground) and the human orthologs of the remaining CG-DMRs (background), separately. We observed q_c SNPs associated with trait c are overlapped with distal feDMRs and the total number of SNPs overlapped with foreground is $Q = \sum_c q_c$. Similarly, $B = \sum_c b_c$ SNPs are overlapped with background and b_c of them are the associated with trait c. For each trait c, the null hypothesis is that q_c follows a hypergeometric distribution, with population size $N = Q + B$, $n_c = q_c + b_c$ are the number of successes (here the number of SNPs related to c that are overlapped with either foreground or background) and sampling number Q. Let X be a random variable representing the observed number of SNPs that are related to trait c and overlapped with foreground. Thus, the probability of observing X is equal to or greater than q_c (i.e.

p-value) is calculated as:

$$P(Xq_c|N, q_c + b_c, Q, c) = \sum_{x=q_c}^{Q} \frac{\binom{n_c}{x}\binom{N-n_c}{Q-x}}{\binom{N}{Q}} \tag{4.2}$$

Using this statistical approach, for SNPs associated with each trait, we tested for their enrichment in the human orthologs of distal feDMRs compared to background. We then used Benjamini-Hochberg approach to adjust the p-values for multiple testing. P-value cutoff given 5% false discovery rate (FDR) was used to call significant enrichment. This procedure was conducted separately for the distal feDMRs of each tissue/organ.

4.7.19 Categorizing CG-DMRs

To understand the potential function of CG-DMRs, we grouped them into various categories based on their genomic location and chromatin signatures. First, we overlapped CG-DMRs with promoters, CGIs and CGI shores and define the overlapping CG-DMRs as proximal CG-DMRs. Out of the 153,019 proximal CG-DMRs, 46,692, 90,831, 1,710 and 13,786 are overlapped with CGI promoters, non-CGI promoters, CGIs and CGI shores, respectively. We avoided assigning proximal CG-DMRs into multiple categories by prioritizing the four genomic features as CGI promoter, non-CGI promoter, CGI and CGI shores (decreasing priority). CG-DMRs were assigned to the category with highest priority. The remaining 1,655,791 CG-DMRs (termed distal CG-DMRs) were further grouped. 415,227 of them were predicted as feDMRs and we called them distal feDMRs. Please note that proximal CG-DMRs also contain feDMRs. (1) Next, we found 221,960 CG-

DMR are on the flanking regions of distal feDMRs and we called them flanking distal feDMRs or fd-feDMRs. fd-feDMRs have no overlap with proximal CG-DMRs or feDMRs. (2) Out of the remaining unclassified CG-DMRs, 194,610 showed strong tissue-specific hypoemethylation pattern and could be assigned to tissue types (tissue-specific CG-DMRs). We called these CG-DMRs as primed distal feDMRs because they showed enrichment of H3K4me1 but not other profiled histone marks in the tissue where they are hypomethylated. (3) The remaining CG-DMRs were defined as unexplained CG-DMRs (unxDMRs). We further divided unxDMRs into two classes by overlapping them with transposable elements: te-unxDMR (transposable element overlapping unxDMRs) and nte-unxDMR (transposable element non-overlapping unxDMRs).

4.7.20 Evolutionary Conservation of CG-DMRs

The evolutionary conservation of CG-DMRs were measured using phyloP score[65]. We first downloaded phyloP score from UCSC genome browser[52] (http://hgdownload.cse.ucsc.edu/goldenpath/mm10/phyloP60way/ mm10.60way.phyloP60way.bw). Next, Deeptools[56] was used to generate the profile of evolutionary conservation of the CG-DMR centers and +/- 5kb flanking regions using options reference-point –referencePoint=center -a 5000 -b 5000.

4.7.21 Finding TF-binding motif enriched in flanking distal feDMRs

To identify the TF-binding motifs enriched in fd-feDMRs relative to feDMRs, we performed motif analysis using the former as foreground and the latter as background. Specifically, for each tissue, the tissue-specific feDMRs were used as background, while the fd-feDMRs that were within 1kb to these tissue-specific feDMRs were used as foreground. Both foreground and background were extended from both sides such that both had mean size 400bp to avoid potential bias residing in different size distribution. Next, hypergeometric test was performed to find TF-binding motifs that were significantly enriched in foreground. The test is the same as the test used in the identification of TF-binding motifs in feDMRs.

4.7.22 TF-binding motif enrichment analysis on primed distal feDMRs

We also performed motif analysis to find TF-binding motifs enriched in pd-feDMRs. The procedure is similar to the motif enrichment analysis on feDMRs. For each tissue, the pd-feDMRs hypomethylated in that tissue were foreground while the remaining pd-feDMRs were background. Then, hypergeometric test was performed to identify significant motif enrichment. Next, for each tissue type, we compared the TF-binding motifs enriched in pd-feDMRs and the tissue-specific feDMRs. Hypergeometric test was used to test the significance of overlap the chance of getting the observed overlap if the two lists were based on random sampling (without replacement) from the TF-binding motifs with TF expression level greater than 10 TPM.

4.7.23 Permutation test to check the overlap between unxDMR and TEs

To estimate the significance of overlap between unxDMRs and TEs, we shuffled the location of unxDMRs using shuffleBed tool from bedtools[61] with default setting and recalculated the overlaps. After repeating this step for 1,000 times, we got an empirical estimate of the overlap if unxDMRs were randomly distributed in the genome. Let the observed number of TE overlapping unxDMRs be x^{obs} and the number of TE overlapping shuffled unxDMRs in permutation i be x_i^{permut}. Lastly, we calculated p-values as

$$p = \frac{(\sum_{i=1}^{1000} I(x^{obs} \leq x_i^{permut})) + 1}{1000 + 1} \tag{4.3}$$

where $I(x) = \begin{cases} 1 & x\ is\ true \\ 0 & x\ is\ false \end{cases}$.

4.7.24 Quantification of the mCG dynamics in tissue-specific CG-DMRs

To quantify the mCG dynamics, we defined and counted loss-of-mCG and gain-of-mCG events. A loss-of-mCG (Gain-of-mCG) event is a decrease (increase) of mCG level by at least 0.1 in one CG-DMR in one stage interval. For example, if the mCG level of one CG-DMR at E11.5 and E12.5 is 0.8 and 0.7 in heart respectively, it is a loss-of-mCG event occurred on the stage interval E11.5-E12.5. Stage interval is defined as the transition between two sampled adjacent stages, e.g. E15.5 and E16.5.

4.7.25 Clustering Forebrain-specific CG-DMRs based on mCG and H3K27ac dynamics

We used k-means clustering to identify subgroups of forebrain-specific CG-DMR based on mCG and H3K27ac dynamics. First, for each forebrain-specific CG-DMR, we calculated the mCG level and H3K27ac enrichment in forebrain samples from E10.5 to adult stages. Here, we used the methylome data of postnatal 1, 2 and 6 week frontal cortex from Lister et al[9] to approximate the DNA methylation landscape of adult forebrain. We also incorporated the H3K27ac data of postnatal 1, 3 and 7 week forebrain samples. Next, to make the range H3K27ac enrichment values comparable to that of mCH levels, for each forebrain-specific CG-DMR, the negative H3K27ac enrichment values were thresholded as zero and we then divided each value by the maximum. If the maximum was zero for some forebrain-specific CG-DMRs, we set all values to be zero. Last, k-means clustering was used to group forebrain-specific CG-DMRs into 8 subgroups. We tried to identify more subgroups but no new patterns were found. Lastly, we used GREAT[66] with Single nearest gene association strategy to find the enriched gene ontology terms of genes near CG-DMRs in each subgroup.

4.7.26 Association between mCG level and H3K27ac enrichment

To investigate the association between mCG and H3K27ac, for each tissue and each developmental stage, we first divided tissue-specific CG-DMRs into three categories: H (highly CG methylated; mCG level ¿ 0.6), M (moderately CG methylated; 0.2 ¡ mCG level $\leq$ 0.6) and L (lowly CG methylated; mCG level $\leq$ 0.2). Then, we checked the distribution of H3K27ac enrichment in different groups of CG-DMRs. To do that, we counted the number of CG-DMRs showing each of the four levels of H3K27ac: $[0, 2], (2, 4], (4, 6]$ and $(6, \infty)$.

4.7.27 Weighted correlation network analysis (WGCNA)

We used weighted correlation network analysis (WGCNA)[46], an unsupervised method, to detect sets of genes with similar expression profiles across samples (R package, WGCNA version 1.51). Briefly, TPM values were First log2 transformed (with pseudo count 1e-5). Then, the TPM values of every gene across all samples were compared against the expression profile of all other genes and a correlation matrix is obtained. To obtain connection strengths between any two genes, we transformed this matrix to an adjacency matrix using a power adjacency function. To choose the parameter (soft threshold) of the power adjacency function, we used the scale-free topology (SFT) criterion, where the constructed network is required to at least approximate scale-free topology. The SFT criterion recommends use of the first threshold parameter value where model-fit saturation is reached as long as it is above 0.8. In this study, the threshold was reached for a power of 5. Next, the adjacency matrix is further transformed to a topological overlap matrix (TOM) that finds neighborhoods of every gene iteratively, based on the connection strengths. The TOM was calculated based on the adjacency matrix derived using the signed hybrid network type, biweight mid correlation and signed TOMtype parameters of the TOMsimilarityFromExpr module in WGCNA. Hierarchical clustering of the TOM was done using the flashClust module using the average method. Then, We used the cutreeDynamic module with the hybrid method, deepSplit = 3 and minClusterSize = 30 parameters to identify modules that have at least 30 genes. A summarized module-specific expression profile is created using the expression of genes within the given module, represented by the eigengene. The eigengene is defined as the first principal component of the log2 transformed TPM values of all genes in a module. In other words, this is a virtual gene that represents the expression profile of all genes in a given module. Next, very similar modules are merged after a hierarchical clustering of the eigengenes of all modules applying a distance threshold of 0.15. Last, the eigengenes are recalculated for all modules after merging.

4.7.28 Gene Ontology Analysis of Genes in each co-expression module (CEM)

To understand the biological meaning of genes in each CEM, we used Enrichr[67, 68] (http://amp.pharm.mssm.edu/Enrichr/) to identify the enriched gene ontology terms in the GO_Biological_Process_2015 category.

4.7.29 Correlating eigengene expression with mCG and enhancer score of feDMRs

We investigated the association between gene expression and epigenomic signatures of regulatory elements in CEMs. First, for each CEM, we used the eigengene expression to summarize the transcription patterns of all genes in the module. Then, we calculated the normalized average enhancer score and normalized average mCG level of all feDMRs that were linked to the genes in the CEM. Specifically, to reduce the potential batch effect, for each tissue and each stage, we normalized the enhancer score of each feDMR by the mean enhancer score of all feDMRs. mCG levels of feDMRs were normalized in similar way except that the data of all DMRs was used to calculate the mean mCG level for each tissue and each stage. Next, for each CEM, the TPM of eigengene, the normalized average enhancer score and mCG level of linked feDMRs were, respectively, further converted to z-scores across stages for each tissue type (in analysis for tissue-specific expression) or across tissue types for each development stage (in analysis for temporal expression). Lastly, for each CEM, we calculated pearson correlation coefficient (R 3.3.1) between the z-score of eigengene expression and the z-score of normalized enhancer score (or mCG level) for each module. The correlation coefficients were calculated in two different settings: 1) for each tissue type, the correlation was computed on z-score of normalized eigengene expression values and enhancer scores (or mCG levels) across different development stages or 2) for each developmental stage, the correlation was computed across different tissue types. The coefficients from former analysis indicate how well temporal gene expression is correlated with enhancer score or mCG level of regulatory elements, while the latter measure the association for tissue-specific gene expression. We then test whether the correlation we observed was significant by comparing it with the correlation based on shuffle data. In the analysis for tissue-specific expression, given a tissue type, we mapped the eigengene expression of one CEM to the enhancer score (or mCG level) of feDMRs linked to the genes in a randomly chosen CEM. For example, in the shuffle setting, when given tissue type was heart, we calculated the correlation between the eigengene expression of CEM14 and the enhancer score of the feDMRs linked to genes in CEM6. In the analysis for temporal expression, given a developmental stage, we performed similar permutation. Next, we calculated the pearson correlation coefficients on this permutation setting. Lastly, using a two-tailed Mann-Whitney test, we compared the median of observed correlation coefficients and the median of those based on shuffled data.

4.7.30 Adult vestigial enhancers and adult active enhancers

The list of adult vestigial enhancer (AD-V enhancer) and adult active enhancer (AD-A enhancers) calls was downloaded from the Supplemental Table 3 in Hon et al[11]. The mm9 coordinates of AD-V and AD-A enhancers were mapped to mm10 reference using liftOver[52] (using default parameters). Only the AD-V and AD-A enhancer calls of heart, intestine and kidney were included because in our dataset, only these three adult tissue samples had matched fetal tissue samples. To trace the enhancer activity of AD-V and AD-A enhancers in fetal tissues, we overlapped AD-V enhancers and AD-A enhancers, respectively, with the tissue-specific feDMRs of the matched tissue type (e.g. we overlapped heart AD-V enhancers with heart-specific feDMRs). intersectBed in bedtools[61] was used to accomplish this analysis and we considered one AD-V/AD-A enhancer to be overlapped with feDMRs if it had at least 1bp within tissue-specific feDMRs.

4.7.31 Dynamic epigenetic modifications in AD-V and AD-A CG-DMRs

To trace the epigenomic changes in AD-V/AD-A enhancers, we first overlapped them with CG-DMRs, which were smaller genomic units for epigenomic changes. CG-DMRs that were overlapped (by at least 1bp) with AD-V (AD-A) enhancers were defined as AD-V (AD-A) CG-DMRs. In each tissue type, the AD-V (AD-A) CG-DMRs that were predicted as tissue-specific feDMRs were AD-V (AD-A) feDMRs whereas the rest were AD-V (AD-A) non-feDMRs.

We calculated the entropy for each AD-V/AD-A feDMR to evaluate the degree of dynamics of enhancer activity of AD-V and AD-A feDMRs as well as the expression of their inferred target genes. Specifically, for tissue type s, we defined $x_{i,s}$ as the value of a metric of feDMRs on developmental stage i, while the metric could be the enhancer score of the feDMR or the $log_{10}(TPM+1)$ value of gene linked to the feDMR. Then, we calculated entropy across developmental stages as

$$entropy = -\sum_{i} p_{i,s} * ln(p_{i,s}) \tag{4.4}$$

where $p_{i,s} = \frac{x_{i,s}}{\sum_{t} x_{t,s}}$.

4.7.32 Identification of CG-DMRs that are marked by H3K27me3 in fetal stages and become AD-V enhancers

In heart, intestine and kidney tissues, we quantified the fraction of tissue-specific CG-DMRs that are enriched for H3K27me3 in fetal stages and show adult vestigial enhancer like epigenomic state (i.e. they escape the remethylation in adult stage and are depleted of H3K4me1 and H3K27ac in adult tissue). Specifically, for each of the three tissues, we selected tissue-specific CG-DMRs whose 1) H3K27me3 signal (normalized RPKM) is greater than 0.5 in at least one fetal stage(s), 2) both H3K4me1 and H3K27ac signals (normalized RPKM) are less than 0.5 in adult stage, and 3) mCG level in adult stage increase by less than 0.1 compared to mCG level in P0. We identified 13,948, 10,757 and 11,268 such CG-DMRs in heart, intestine and kidney respectively

4.7.33 Partially methylated domain (PMD) identification

PMDs were identified as previously described[8]. Briefly, we trained a random forest classifier. To get data used to train the classifier, we first visually selected regions on chromosome 19 that we felt were strong candidates as PMDs or non-PMDs in E14.5 liver sample. We picked 5 PMDs (chr19:46110000-46240000, chr19:45820000-45960000, chr19:47140000-47340000 and chr19:48060000-52910000) and 7 Non-PMD regions (chr19:4713800-4928700, chr19:7420700-7541100, chr19:8738100-8967000, chr19:18633300-18713800, chr19:53315500-53390000, chr19:55256600-55633900 and chr19:59281600-59329200).

Next, these regions were divided into 10kb non-overlapping bins and we calculated the percentiles of the methylation levels at the CG sites within each bin. CG sites that are within CGIs, DMVs[69] or any of four Hox loci (see below) were excluded because they are typically hypomethylated and may result in incorrect PMD calling. Sites with less than 5 reads covered were not considered either. We trained the random forest classifier using data in E14.5 liver (with data of two replicates combined) and we then used it to predict whether a 10kb bin is PMD or non-PMD in all liver samples (with replicates separated). We chose a large bin size (10 kb) to reduce the effect of methylation variations in smaller scale (such as DMRs) as PMDs were first discovered as large (mean length = 153kb, PMID: 19829295) regions with intermediate methylation level (¡ 70%, PMID: 19829295). Furthermore, the features (methylation level distribution of CG sites) used in the classifier required enough CG sites within each bin to robustly estimate the distribution, which necessitated a relatively large bin. Also, we excluded any 10kb bins containing less than 10 CG sites due to the same reason. These percentiles were used as features for the random forest. The random forest implement was from scikit-learn (version 0.17.1)[70] python module and the following arguments were supplied to the Python function RandomForestClassifier from scikit-learn: n_estimators = 10000, max_features=None, oob_score=True, compute_importances=True.

Lastly, we merged consecutive 10kb bins that were predicted as PMD into blocks and filtered out blocks smaller than 100kb. We further excluded blocks that were overlapped with gaps in mm10 genome (downloaded from UCSC genome browser, Sep 21, 2013). To get the PMDs that were reproducible in both replicates, we only considered the genomic regions that were larger than 100kb and were covered by PMD calls in both replicates. These regions were the final PMDs used for later analyses. Because there is only one replicated for adult liver, we skipped this step.

PMDs were originally called without excluding CG sites in Hox regions. We found that four PMDs turned out to be four Hox loci in the mouse genome. Because the Hox loci are more likely to be large DMVs[69], we removed any PMDs that overlap with these four Hox loci (chr11:96257739-96358516, chr15:102896908-103038064, chr2:74648392-74748841 and chr6:52146273-52277140).

4.7.34 Overlapping PMDs with LADs

We downloaded the lamina associated domains (LADs) of normal mouse liver cells (AML12 hepatocyte) from the Table S2 of Fu et al[71]. The mm9 coordinate of LADs was converted to mm10 using liftOver with default settings. We used permutation to test the significance of overlap between PMDs and LADs. Similar to the procedure for checking the overlap between TEs and unxDMRs, we permutated the genomic locations of PMDs for 1,000 times and recorded the number of overlapping bases (x_i^{shuf} for permutation i) between shuffled PMDs and LADs. Then, we compared x_i^{shuf} with the observed numbers of overlapping bases (x^{obs}) between PMDs and LADs and computed p-values as:

$$p = \frac{(\sum_{i=1}^{1000} I(x^{obs} \leq x_i^{shuf})) + 1}{1000 + 1} \tag{4.5}$$

where $I(x) = \begin{cases} 1 & x\ is\ true \\ 0 & x\ is\ false \end{cases}$.

4.7.35 Replication Timing Data

Replication timing data (build mm10) of three mouse cell types was downloaded from ReplicationDomain[72]. The cell types used for the analysis are mESC (id: 1967902&4177902_TT2ESMockCGHRT), neural progenitor cells (id: 4180202&4181802_TT2NSMockCGHRT) and mouse embryonic fibroblasts (id: 304067-1 Tc1A).

4.7.36 Gene Transcription in PMDs

We obtained PMD overlapping protein-coding genes using intersectBed. Similar approach was used to get the protein-coding genes overlapped with PMD flanking regions (upstream 100kb and downstream 100kb of PMDs) and genes overlapped with PMDs were removed from this list. Lastly, we compared the expression of PMD-overlappign genes (n=5,748) and the genes (n=2,555) overlapped with flanking regions.

4.7.37 Sequence preference of mCH

To interrogate the sequence preference of mCH, as previous described[8], we first identified CH sites that showed significantly higher methylation level than noise due to sodium bisulfite non-conversion. For a CH site, we counted the number of reads that supported methylation and the number of reads that did not.

Next, we performed a binomial test with the success probability equal to the sodium bisulfite non-conversion rate. FDR (1%) was controlled using benjamini-hochberg approach and this analysis was done for each three nucleotide context independently (e.g., a pvalue cutoff was calculated for CAG cytosines). Last, we counted sequence motif occurrence of +/-5bp around the tri-nucleotide context of methylated mCH sites and visualized the sequence preference using seqLogo[73]

4.7.38 mCH domain calling

We used an iterative process to call mCH domains, which are genomic regions that are enriched for mCH compared to flanking regions. First, we selected a set of samples that showed no evidence of mCH. Data of these samples were used in following steps to filter out genomic regions that are prone to misalignment and showed suspicious abundant mCH. Analysis on global mCH level and mCH motif revealed that E10.5 and E11.5 samples excluding heart samples have extremely low mCH and the significantly methylated non-CG sites showed little CA preference. Therefore, we assumed they contain no mCH domain and any mCH domains called by algorithm are likely artifacts. We will show that by filtering out the domains called in the control samples, we were able to exclude the genomic regions that were prone to mapping error or avoid other potential drawbacks in the processing pipeline.

We applied change point detection algorithm on mCH level of 5kb non-overlapping bins across the genome to identify loci where sharp change in mCH levels occurred. We only included bins that contain at a minimum 500 CH sites and at least 50% of CH sites were covered by 10 or more reads. The identified loci are boundaries that separate mCH domains from genomic regions showing background level mCH. We implemented this step using function cpt.mean in R package changepoint, with options method=”PELT”, pen.value=0.05, penalty= ”Asymptotic” and minseglen=2. To match the range of , we scaled up mCH levels by a factor of 1,000.

The iterative procedure was run as follow: 1) Create an empty list of excluded regions. 2) For each control sample, apply change point detection algorithm to the scaled mCH levels of 5kb non-overlapping bins. Bins overlapped with excluded regions were ignored. 3) Segment genome into chunks based on identified change points. 4) Calculate the mCH level of each chunk as the mean mCH level of the overlapping 5kb bins that were not overlapped with excluded regions. 5) Identify mCH domains as chunks whose mCH level was at least 50% greater than the mCH level of both upstream and downstream chunks. Pseudo mCH level 0.001 was used to avoid dividing zero. 6) Add mCH domains to the list of excluded regions. 7) Repeat step 2 to 6 until the list of excluded regions stop expanding. 8) Apply step 2-5 to all samples. 9) For each tissue/organ, only retain the regions that are identified as (part of) mCH domain in both replicates and filter out any that are less than 15kb in length they need to span at least three bins. These are the mCH domains of that tissue/organ. 10) Merge the mCH domains in all tissues and organs to get a list of combined mCH domains.

4.7.39 Clustering mCH Domains

We applied k-means clustering to group the 384 mCH domains into 5 clusters based on the normalized mCH accumulation profile of each mCH domain and corresponding flanking regions (100kb upstream and 100kb downstream). Specifically, 1) in each tissue sample, the mCH accumulation profile of one mCH domain was represented as a vector of length 50: the mCH level of 20 5kb bins upstream mCH domain, 10 bins that equally divided mCH domain and 20 5kb bins downstream. 2) Then, we normalized all values by the average mCH level of bins of flanking regions (the 20 5kb bins upstream and 20 5kb bins downstream of mCH domain). 3) We computed the profile in samples of 6 tissue types (midbrain, hindbrain, heart, intestine, stomach and kidney) that showed the most evident mCH accumulation in fetal development. 4) Using the profile in these tissue samples, k-means (R v3.3.1) was used to clustered mCH domains with k = 5. We also tried higher cluster numbers (e.g. 6) but found not new pattern. Even in current setting (k=5), the mCH domains in cluster 1 (C1) and cluster 3 (C3) shared similar mCH accumulation pattern.

4.7.40 Genes in mCH domains

We obtained the overlapping genes of mCH domains by overlapping gene bodies with mCH domains using intersectBed in bedtools[61]. Only protein coding genes were considered and we further filtered out any genes with names starting with Rik or Gm[0-9], where [0-9] represents a single digit. For the overlapping genes of each mCH domain cluster, we used EnrichR[67, 68] to find the enriched gene ontology terms (GO_Biological_Process_2015).

Next we asked whether the overlapping genes were enriched for TF encoding genes. List of mouse TFs was downloaded from AnimalTFDB[74] (Feb 27, 2017). Then, we performed permutation test to estimate the significance. Specifically,

x^{obs} is the number of TF encoding genes in all overlapping genes. We randomly selected the same number of genes for 1000 times and in the ith time, x_i^{permut} of the randomly selected genes encoded TF. Last, p-values was calculated as

$$p = \frac{(\sum_{i=1}^{1000} I(x^{obs} \leq x_i^{permu})) + 1}{1000 + 1} \tag{4.6}$$

where $I(x) = \begin{cases} 1 & x\ is\ true \\ 0 & x\ is\ false \end{cases}.$

4.7.41 mCH accumulation indicates gene repression

To evaluate the association between mCH abundance and gene transcription, we traced the expression dynamics of genes inside mCH domains. For mCH domains in each cluster, we first calculated the TPM z-score for each of the overlapping genes. Specifically, for each tissue type and each overlapping gene, we normalized TPM values in the samples of that tissue type to z-scores. The z-scores showed the trajectory of dynamic expression, in which the aptitude information of expression was removed. If the gene was not expressed, we did not perform the normalization. Next, we calculated the z-scores for all genes that have no overlapped with any mCH domains. Lastly, we subtracted the z-scores of overlapping genes by the z-scores of all genes outside mCH domains. The resulting values indicated how genes in mCH domains were regulated differently relative to genes not in mCH domains.

4.7.42 feDMRs in mCH domains

To evaluate whether feDMRs were enriched in mCH domains, we calculated the percentage of bases within mCH domains that were also within tissue-specific feDMRs. Specifically, we first divided the genome into non-overlapping 100bp bins and then, for each tissue, we calculated the percentage of bases in each bin that were overlapped with tissue-specific feDMRs. Next, we plotted the percentages across mCH domains (each was equally divided into 10 non-overlapping bins) and flanking regions (10kb upstream and 10kb downstream; each contained 10 1kb bins). For the feDMRs that were overlapped with mCH domains, we then traced their mCG level changed over development. For each tissue type and each developmental stage, we calculated the percentage of tissue-specific feDMRs whose mCG level was in each range: $[0 - 0.2), [0.2, 0.4), [0.4, 0.6), [0.6, 0.8)$ and $[0.8, 1]$. For each tissue and each developmental stage, the percentages were calculated for tissue-specific feDMRs in mCH domains and also for all tissue-specific feDMRs. Last, the ratio of the former to the later was defined as the enrichment of feDMRs (with certain mCG level) in mCH domains, which was then transformed to log2 ratio.

4.8 Figures

Figure 4.1: **Annotation of methylation variable regulatory elements in developing mouse tissues.** **a,** Tissue samples (green) profiled in this study. Grey cells mark the tissues and stages that were not sampled because either the tissue is not yet formed or it was not possible to obtain enough material for experiment, or the tissue-type was heterogeneous to obtain informative data. **b,** Genome-wide CG methylation (mCG) levels of each tissue across their developmental trajectories. The data point of adult forebrain (postnatal 6 week frontal cortex) is from Lister et al[9]. **c,** An example of a CG differentially methylated region (CG-DMR) in the body of Satb2 gene. Top two tracks show the gene annotation and the locations of CG islands (CGIs), which are followed by mCG tracks and one CG-DMR track. Gold ticks represent methylated CG sites and their heights indicate the mCG level, ranging from 0 to 1. Ticks on the forward DNA strand are projected upward and ticks on the reverse DNA strand are projected downward. **d,** Fetal CG-DMRs identified in this study cover majority of the adults CG-DMRs from a previous study of adult tissues[36]. The numbers related to fetal CG-DMR in this study are shown without parenthesis, whereas the numbers in parenthesis are related to adult tissue CG-DMRs. **e,** Distance of CG-DMRs to the nearest transcription start sites (TSSs). **f,** Categorization of CG-DMRs. proximal CG-DMRs are CG-DMRs that are overlapped with promoters, CG islands (CGIs) or CGI shores. The remaining CG-DMRs are defined as distal CG-DMRs. fetal enhancer-linked CG-DMRs (feDMRs) are those predicted to show enhancer activity using REPTILE algorithm[32], which contain 415,227 distal feDMRs and 72,140 proximal feDMRs. CG-DMRs within 1kb to distal feDMRs are flanking distal CG-DMRs. Of the remaining distal CG-DMRs, we defined primed distal feDMRs as those showing primed enhancer-like chromatin signatures. The remaining CG-DMRs are unexplained distal CG-DMRs (unxDMRs), whose functions are unknown. unxDMRs are further stratified based on their overlap with transposons: transposal element overlapping unxDMRs (te-unxDMRs) and transposal element non-overlapping unxDMRs (nte-unxDMRs). The number of CG-DMRs assigned to each group is shown in the parentheses. See Methods for details. **g,** Conservation (phyloP) score of promoters and different categories of distal CG-DMRs. **h,** Tissue-specific feDMRs are enriched in GWAS SNPs associated with tissue/organ specific functions and tissue-related disease states.

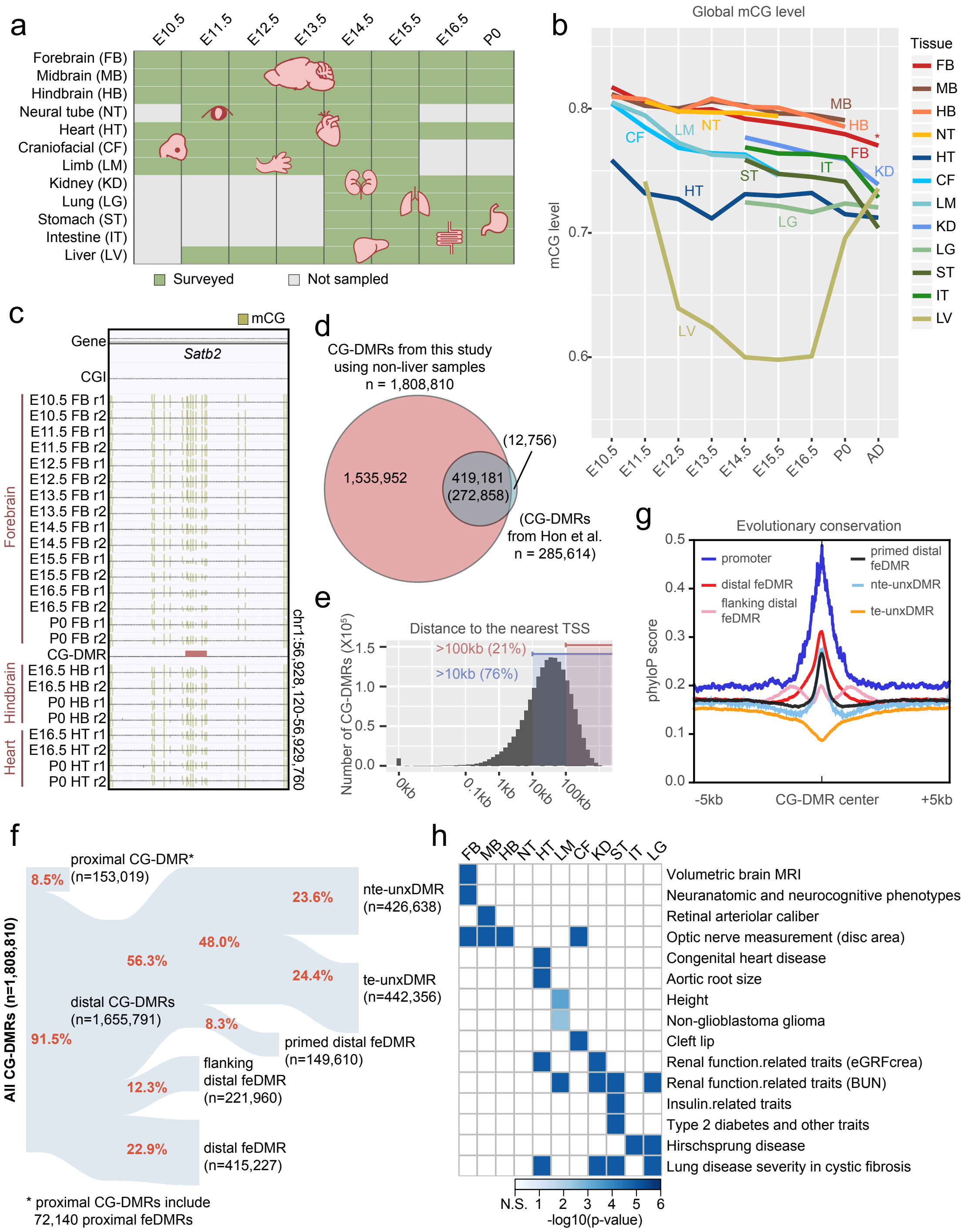

a
E10.5
E11.5
E12.5
E13.5
E14.5
E15.5
E16.5
P0
Forebrain (FB)
Midbrain (MB)
Hindbrain (HB)
Neural tube (NT)
Heart (HT)
Craniofacial (CF)
Limb (LM)
Kidney (KD)
Lung (LG)
Stomach (ST)
Intestine (IT)
Liver (LV)
Surveyed
Not sampled
b
Global mCG level
mCG level
0.8
0.7
0.6
MB
HB
FB
NT
LM
CF
ST
IT
KD
HT
LG
LV
E10.5
E11.5
E12.5
E13.5
E14.5
E15.5
E16.5
P0
AD
Tissue
FB
MB
HB
NT
HT
CF
LM
KD
LG
ST
IT
LV
c
mCG
Gene
Satb2
CGI
Forebrain
E10.5 FB r1
E10.5 FB r2
E11.5 FB r1
E11.5 FB r2
E12.5 FB r1
E12.5 FB r2
E13.5 FB r1
E13.5 FB r2
E14.5 FB r1
E14.5 FB r2
E15.5 FB r1
E15.5 FB r2
E16.5 FB r1
E16.5 FB r2
P0 FB r1
P0 FB r2
CG-DMR
Hindbrain
E16.5 HB r1
E16.5 HB r2
P0 HB r1
P0 HB r2
Heart
E16.5 HT r1
E16.5 HT r2
P0 HT r1
P0 HT r2
chr1:56,928,120-56,929,760
d
CG-DMRs from this study using non-liver samples n = 1,808,810
1,535,952
419,181 (272,858)
(12,756)
(CG-DMRs from Hon et al. n = 285,614)
e
Distance to the nearest TSS
Number of CG-DMRs (X10^5)
1.5
1.0
0.5
0.0
>100kb (21%)
>10kb (76%)
0kb
0.1kb
1kb
10kb
100kb
g
Evolutionary conservation
phyloP score
0.5
0.4
0.3
0.2
0.1
0.0
promoter
distal feDMR
flanking distal feDMR
primed distal feDMR
nte-unxDMR
te-unxDMR
-5kb
CG-DMR center
+5kb
f
All CG-DMRs (n=1,808,810)
8.5%
proximal CG-DMR* (n=153,019)
91.5%
56.3%
distal CG-DMRs (n=1,655,791)
48.0%
23.6%
nte-unxDMR (n=426,638)
24.4%
te-unxDMR (n=442,356)
8.3%
primed distal feDMR (n=149,610)
12.3%
flanking distal feDMR (n=221,960)
22.9%
distal feDMR (n=415,227)
* proximal CG-DMRs include 72,140 proximal feDMRs
h
FB
MB
HB
NT
HT
LM
CF
KD
ST
IT
LG
Volumetric brain MRI
Neuranatomic and neurocognitive phenotypes
Retinal arteriolar caliber
Optic nerve measurement (disc area)
Congenital heart disease
Aortic root size
Height
Non-glioblastoma glioma
Cleft lip
Renal function.related traits (eGRFcrea)
Renal function.related traits (BUN)
Insulin.related traits
Type 2 diabetes and other traits
Hirschsprung disease
Lung disease severity in cystic fibrosis
N.S.
1
2
3
4
5
6
-log10(p-value)

Figure 4.2: **Tissue-specific CG-DMRs undergo continuous demethylation during embryogenesis and remethylation after birth. a,** CG methylation (mCG) level of tissue-specific CG-DMRs. The number under each heatmap indicates the number of tissue-specific CG-DMRs. mCG data from adult (AD) forebrain was approximated using data from postnatal 6 week frontal cortex from Lister et al9. **b,** The numbers of loss-of-mCG events (blue) and gain-of-mCG events (red) in tissue-specific CG-DMRs for each fetal stage interval. We defined one loss-of-mCG (gain-of-mCG) event as the decrease (increase) of at least 0.1 in mCG level of one CG-DMR for one fetal stage interval. **c-d,** Fraction of tissue-specific CG-DMRs that undergo lost-of-mCG (blue) and gain-of-mCG (red) during development. The blue (loss-of-mCG) or red (gain-of-mCG) line shows the aggregated values over all non-liver tissues, whereas grey lines show the data for each tissue type. **e,** mCG and H3K27ac dynamics of forebrain-specific CG-DMRs. Frontal cortex methylomes from postnatal 1, 2, 4, 6 weeks (P1w to P6w) were compared with data from adult forebrain. Forebrain-specific CG-DMRs were clustered into 8 groups (see Methods for details). **f,** Relationship between mCG level and enrichment of H3K27ac in tissue-specific CG-DMRs. For each tissue type, tissue-specific CG-DMRs were first grouped into three categories (L: low; M: median; H: high) based on their mCG level. Then, the fraction of tissue-specific CG-DMRs from each category that showed different levels of H3K27ac enrichment was quantified. This panel shows the results of all non-liver tissues.

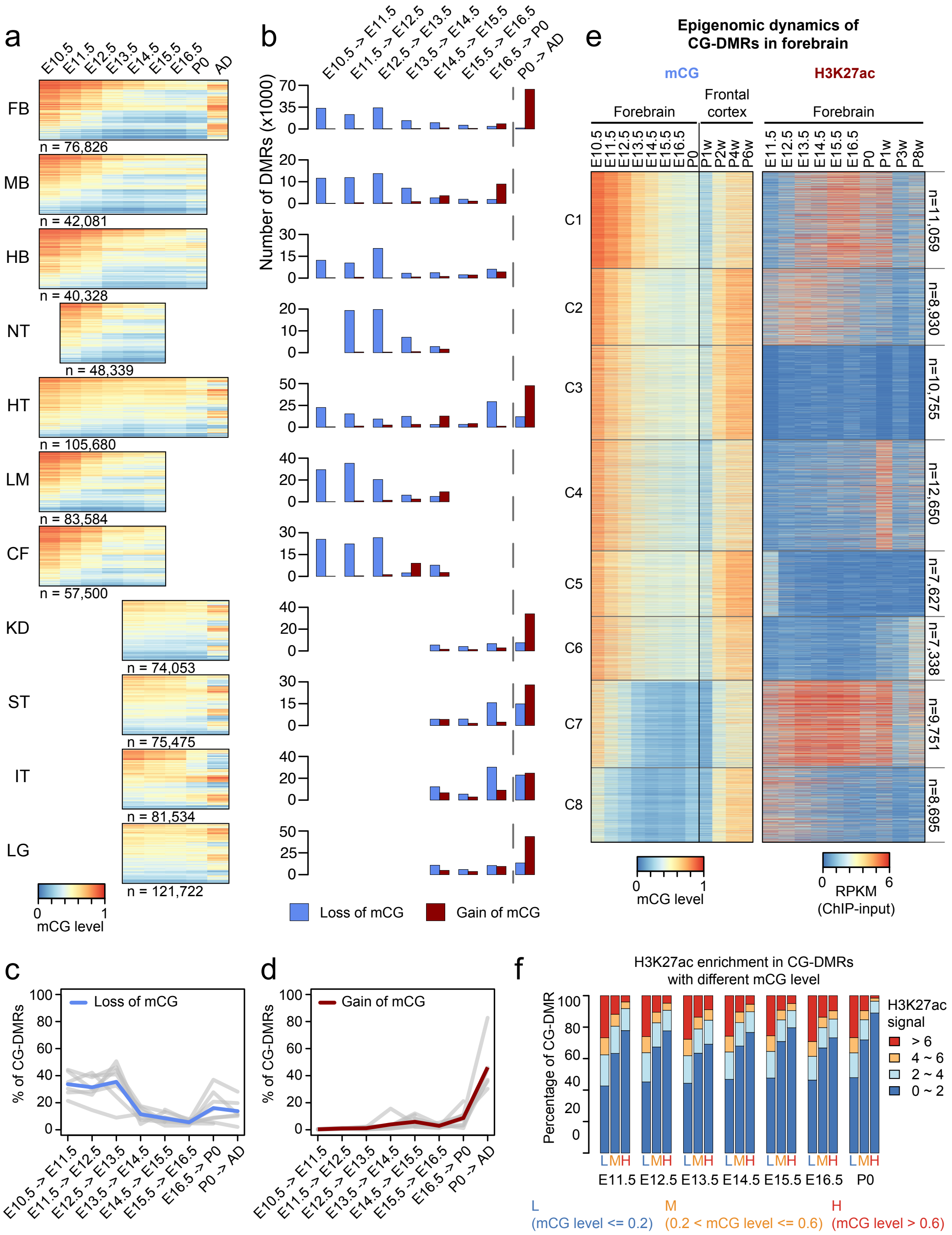

a
E10.5 E11.5 E12.5 E13.5 E14.5 E15.5 E16.5 P0 AD
FB n = 76,826
MB n = 42,081
HB n = 40,328
NT n = 48,339
HT n = 105,680
LM n = 83,584
CF n = 57,500
KD n = 74,053
ST n = 75,475
IT n = 81,534
LG n = 121,722
0 1
mCG level
b
E10.5 -> E11.5
E11.5 -> E12.5
E12.5 -> E13.5
E13.5 -> E14.5
E14.5 -> E15.5
E15.5 -> E16.5
E16.5 -> P0
P0 -> AD
Number of DMRs (x1000)
70 35 0
20 10 0
30 15 0
20 10 0
50 25 0
40 20 0
30 15 0
40 20 0
30 15 0
40 20 0
50 25 0
Loss of mCG
Gain of mCG
c
% of CG-DMRs
100 80 60 40 20 0
Loss of mCG
E10.5 -> E11.5
E11.5 -> E12.5
E12.5 -> E13.5
E13.5 -> E14.5
E14.5 -> E15.5
E15.5 -> E16.5
E16.5 -> P0
P0 -> AD
d
% of CG-DMRs
100 80 60 40 20 0
Gain of mCG
E10.5 -> E11.5
E11.5 -> E12.5
E12.5 -> E13.5
E13.5 -> E14.5
E14.5 -> E15.5
E15.5 -> E16.5
E16.5 -> P0
P0 -> AD
e
Epigenomic dynamics of
CG-DMRs in forebrain
mCG
H3K27ac
Forebrain
Frontal cortex
E10.5 E11.5 E12.5 E13.5 E14.5 E15.5 E16.5 P0
P1w P2w P4w P6w
Forebrain
E11.5 E12.5 E13.5 E14.5 E15.5 E16.5 P0 P1w P3w P8w
C1 n=11,059
C2 n=8,930
C3 n=10,755
C4 n=12,650
C5 n=7,627
C6 n=7,338
C7 n=9,751
C8 n=8,695
0 1
mCG level
0 6
RPKM
(ChIP-input)
f
H3K27ac enrichment in CG-DMRs
with different mCG level
Percentage of CG-DMR
100 80 60 40 20 0
H3K27ac signal
> 6
4 ~ 6
2 ~ 4
0 ~ 2
LMH LMH LMH LMH LMH LMH LMH
E11.5 E12.5 E13.5 E14.5 E15.5 E16.5 P0
L
(mCG level <= 0.2)
M
(0.2 < mCG level <= 0.6)
H
(mCG level > 0.6)

Figure 4.3: **The methylomes and transcriptomes of human tissues. a,** Expression of the 2,500 most variable genes in all tissue samples. Tissue samples were grouped using hierarchical clustering. Gene expression is measured by log10 (TPM+1) (transcripts per million) z-score. Black box highlights a group of co-expressed genes that are highly expressed in neuronal tissues. **b,** 33 CEMs identified in WGCNA and their eigengene expression. The CEMs related to the next panel (c) are bolded. **c,** The most enriched gene ontology (Biological Process) terms of genes in four representative modules. **d,** Correlation of the tissue-specific eigengene expression (orange) for each developmental stage with the average mCG level (blue) and the average enhancer score (red) for feDMRs linked to the genes in the CEM32. The mCG levels (enhancer score) of each feDMR was normalized by dividing the genome-wide average mCG level (enhancer score) and then transformed to z-scores (See Methods for details). Pearson correlation coefficient (r) was calculated. **e,** Pearson correlation coefficients of mCG (blue) or enhancer score (red) of neighboring feDMRs with tissue-specific eigengene expression across all 33 CEMs on all stages. P-values were obtained by testing the median of the Pearson correlation coefficients against the median of the shuffled (grey) using two-tailed Mann-Whitney test. **f-g,** Similar to (d), correlation of temporal eigengene expression for CEM29 (f) and CEM12 (g) with the average mCG level and the average enhancer score of neighboring feDMRs. **h,** Similar to (e), Pearson correlation coefficients of mCG and enhancer score with temporal epigengene expression across all CEMs and all tissue types, excluding liver. See Methods for details.

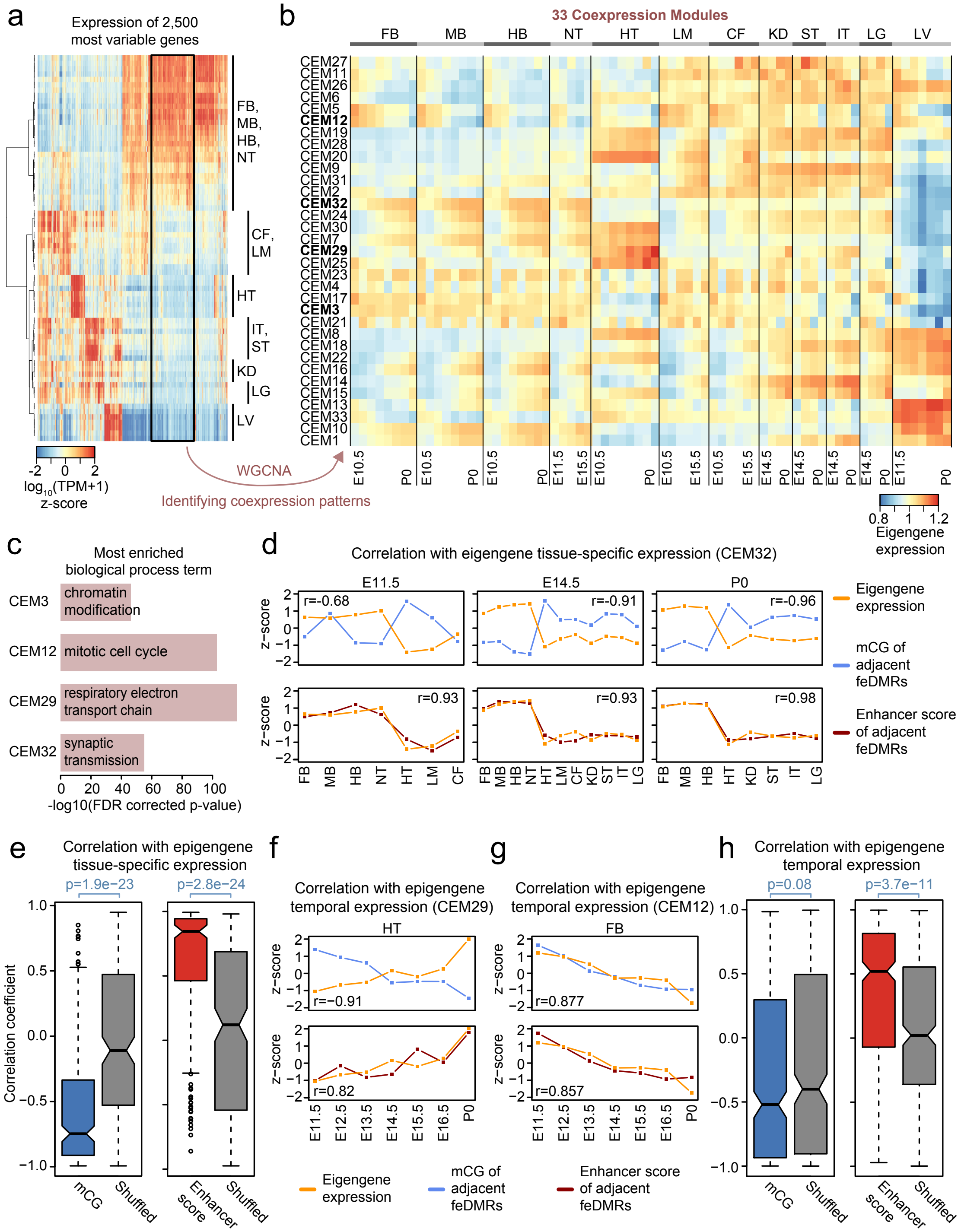

a
Expression of 2,500
most variable genes
FB,
MB,
HB,
NT
CF,
LM
HT
IT,
ST
KD
LG
LV
-2 0 2
log10(TPM+1)
z-score
WGCNA
Identifying coexpression patterns
b
33 Coexpression Modules
FB MB HB NT HT LM CF KD ST IT LG LV
CEM27
CEM11
CEM26
CEM6
CEM5
CEM12
CEM19
CEM28
CEM20
CEM9
CEM31
CEM2
CEM32
CEM24
CEM30
CEM7
CEM29
CEM25
CEM23
CEM4
CEM17
CEM3
CEM21
CEM8
CEM18
CEM22
CEM16
CEM14
CEM15
CEM13
CEM33
CEM10
CEM1
E10.5 P0 E10.5 P0 E10.5 P0 E11.5 E15.5 E10.5 P0 E10.5 E15.5 E10.5 E15.5 E14.5 P0 E14.5 P0 E14.5 P0 E14.5 P0 E11.5 P0
0.8 1 1.2
Eigengene
expression
c
Most enriched
biological process term
CEM3 chromatin modification
CEM12 mitotic cell cycle
CEM29 respiratory electron transport chain
CEM32 synaptic transmission
0 20 40 60 80 100
-log10(FDR corrected p-value)
d
Correlation with eigengene tissue-specific expression (CEM32)
E11.5
E14.5
P0
r=-0.68
r=-0.91
r=-0.96
r=0.93
r=0.93
r=0.98
z-score
FB MB HB NT HT LM CF
FB MB HB NT HT LM CF KD ST IT LG
FB MB HB HT KD ST IT LG
Eigengene expression
mCG of adjacent feDMRs
Enhancer score of adjacent feDMRs
e
Correlation with epigengene
tissue-specific expression
p=1.9e−23
p=2.8e−24
Correlation coefficient
1.0 0.5 0.0 −0.5 −1.0
mCG
Shuffled
Enhancer score
Shuffled
f
Correlation with epigengene
temporal expression (CEM29)
HT
r=−0.91
r=0.82
g
Correlation with epigengene
temporal expression (CEM12)
FB
r=0.877
r=0.857
z-score
E11.5 E12.5 E13.5 E14.5 E15.5 E16.5 P0
Eigengene expression
mCG of adjacent feDMRs
Enhancer score of adjacent feDMRs
h
Correlation with epigengene
temporal expression
p=0.08
p=3.7e−11
1.0 0.5 0.0 −0.5 −1.0
mCG
Shuffled
Enhancer score
Shuffled

Figure 4.4: **Dynamic epigenetic signatures of adult vestigial enhancers.** **a,** A VISTA enhancer browser[37] image showing an example of an adult vestigial enhancer (AD-V enhancer) overlapping an experimentally validated heart enhancers in E11.5 embryo. **b,** Fractions of AD-V and AD-A enhancers overlapping feDMRs. **c,** Dynamic epigenetic signatures of heart AD-V CG-DMRs (top) and AD-A CG-DMRs (bottom). Leftmost bars, (red label) show CG-DMRs that were predicted as feDMR. For each category, four heatmaps are displayed (from left to right) and show the enhancer score, mCG level, H3K27ac enrichment and H3K27me3 enrichment for each CG-DMR. **d,** Number of heart AD-V non-feDMRs. Red indicates heavily CG methylated AD-V non-feDMRs (mCG level ¿ 0.6). **e,** Enrichment of H3K27me3 in heart AD-V non-feDMRs. Those that are heavily CG methylated (red; mCG level ¿ 0.6) verses less methylated CG-DMRs (grey; mCG level ¡= 0.6).

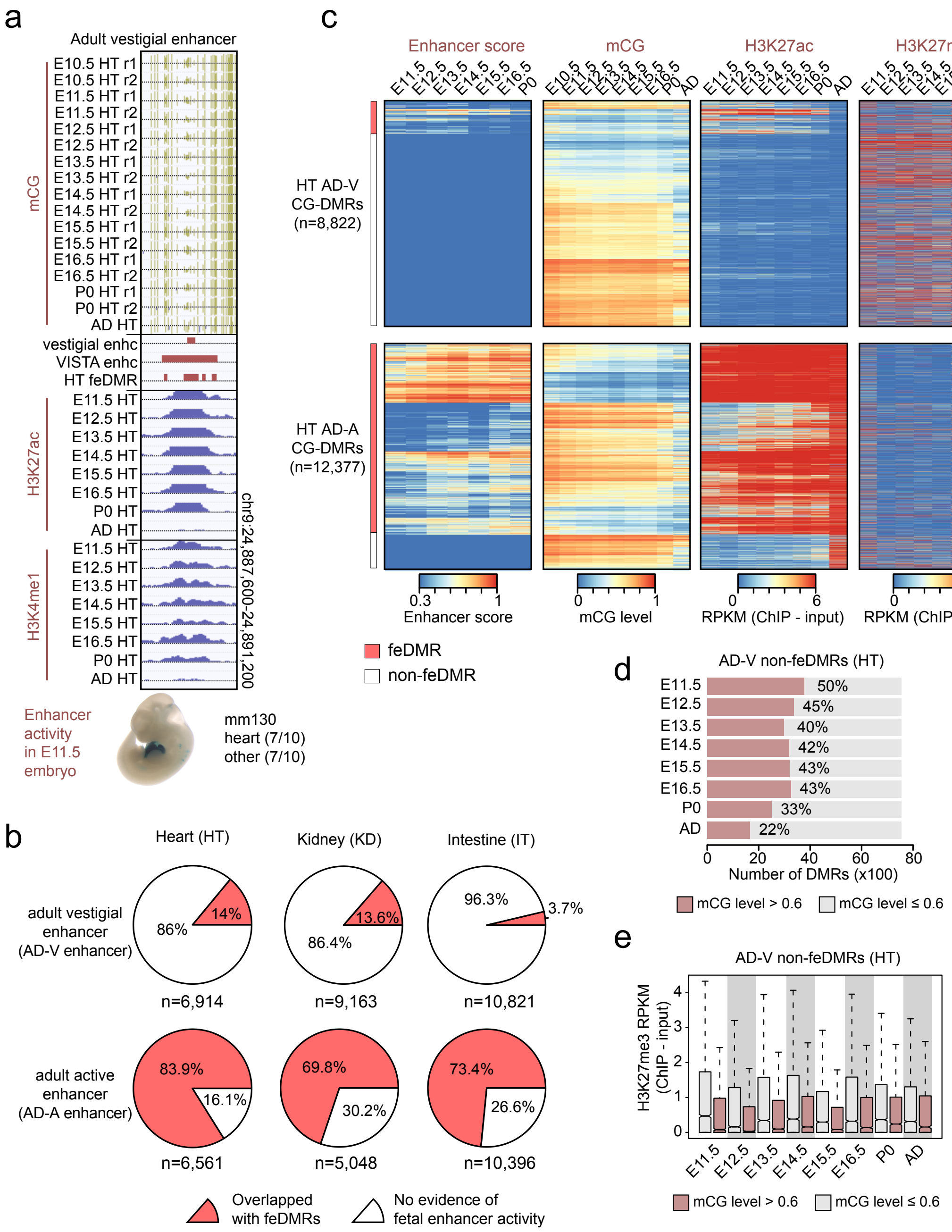

a
Adult vestigial enhancer
mCG
E10.5 HT r1
E10.5 HT r2
E11.5 HT r1
E11.5 HT r2
E12.5 HT r1
E12.5 HT r2
E13.5 HT r1
E13.5 HT r2
E14.5 HT r1
E14.5 HT r2
E15.5 HT r1
E15.5 HT r2
E16.5 HT r1
E16.5 HT r2
P0 HT r1
P0 HT r2
AD HT
vestigial enhc
VISTA enhc
HT feDMR
H3K27ac
E11.5 HT
E12.5 HT
E13.5 HT
E14.5 HT
E15.5 HT
E16.5 HT
P0 HT
AD HT
H3K4me1
E11.5 HT
E12.5 HT
E13.5 HT
E14.5 HT
E15.5 HT
E16.5 HT
P0 HT
AD HT
chr9:24,887,600-24,891,200
Enhancer activity in E11.5 embryo
mm130
heart (7/10)
other (7/10)
b
Heart (HT)
Kidney (KD)
Intestine (IT)
adult vestigial enhancer (AD-V enhancer)
14%
86%
n=6,914
13.6%
86.4%
n=9,163
96.3%
3.7%
n=10,821
adult active enhancer (AD-A enhancer)
83.9%
16.1%
n=6,561
69.8%
30.2%
n=5,048
73.4%
26.6%
n=10,396
Overlapped with feDMRs
No evidence of fetal enhancer activity
c
Enhancer score
mCG
H3K27ac
H3K27me3
E11.5 E12.5 E13.5 E14.5 E15.5 E16.5 P0
E10.5 E11.5 E12.5 E13.5 E14.5 E15.5 E16.5 P0 AD
E11.5 E12.5 E13.5 E14.5 E15.5 E16.5 P0 AD
E11.5 E12.5 E13.5 E14.5 E15.5 E16.5 P0 AD
HT AD-V CG-DMRs (n=8,822)
HT AD-A CG-DMRs (n=12,377)
0.3 1 Enhancer score
0 1 mCG level
0 6 RPKM (ChIP - input)
0 1 RPKM (ChIP - input)
feDMR
non-feDMR
d
AD-V non-feDMRs (HT)
E11.5 50%
E12.5 45%
E13.5 40%
E14.5 42%
E15.5 43%
E16.5 43%
P0 33%
AD 22%
0 20 40 60 80
Number of DMRs (x100)
mCG level > 0.6
mCG level ≤ 0.6
e
AD-V non-feDMRs (HT)
H3K27me3 RPKM (ChIP - input)
0 1 2 3 4
E11.5 E12.5 E13.5 E14.5 E15.5 E16.5 P0 AD
mCG level > 0.6
mCG level ≤ 0.6

Figure 4.5: **mCH accumulation indicates transcriptional repression. a,** Genome-wide non-CG methylation (mCH) levels for each tissue across their developmental trajectories. The adult (AD) forebrain data (postnatal 6 week frontal cortex) is from Lister et al[9]. **b,** An example of a mCH domain. Enriched for mCH accumulation determined by comparison to flanking regions. **c,** K-means clustering identification of 384 mCH domains clustered into 5 groups based on the tissue-specific mCH accumulation. Heatmap showing the methylation profiling of mCH domains and flanking genomic regions (100kb upstream and 100kb downstream). **d,** Number of genes overlapping mCH domains in each of 5 groups. Dark blue bars indicate the number of genes encode transcription factors in mCH domains. Examples of genes located within mCH domains are listed on the right. **e,** The most enriched gene ontology (Biological Process) terms for genes that lie within mCH domains for each cluster. **f,** Expression dynamics of genes within mCH domains relative to the other genes. Z-scores were calculated for each gene across development and each line shows the mean value of mCH overlapping genes for each cluster. **g,** Tissue-specific enrichment of feDMRs in mCH domains. Each mCH domain was divided into 10 bins and its flanking regions included ten 10kb upstream bins and ten 10kb downstream bins. Line plots show the fraction of bases in each bin that are overlapped with tissue-specific feDMRs. A list of tissues in each plot indicates wthere feDMRs are enriched in mCH domains compared to flanking regions (from the most enriched to the least enriched).

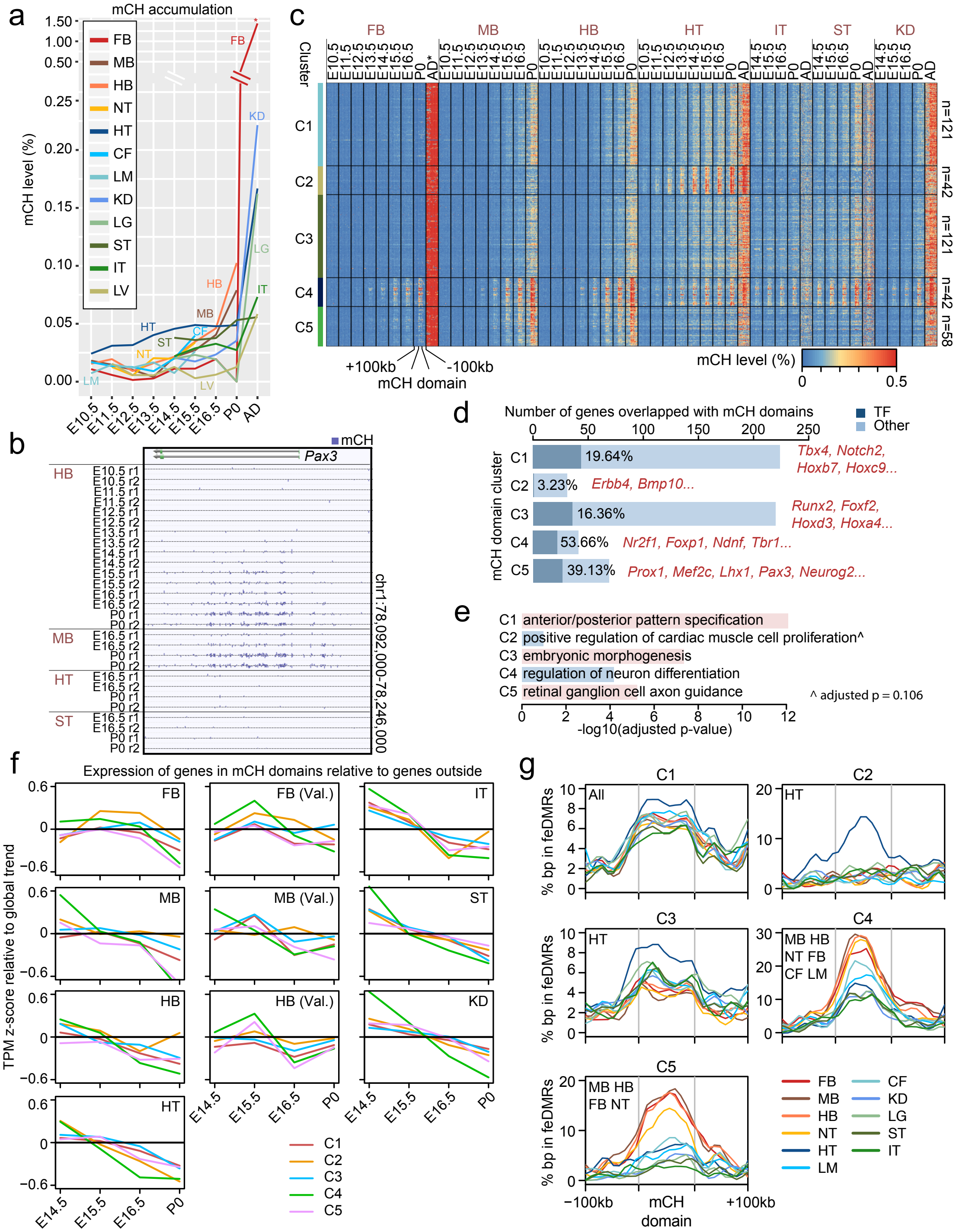

a
mCH accumulation
mCH level (%)
1.50
1.00
0.50
0.25
0.20
0.15
0.10
0.05
0.00
FB
MB
HB
NT
HT
CF
LM
KD
LG
ST
IT
LV
E10.5
E11.5
E12.5
E13.5
E14.5
E15.5
E16.5
P0
AD
b
mCH
Pax3
HB
E10.5 r1
E10.5 r2
E11.5 r1
E11.5 r2
E12.5 r1
E12.5 r2
E13.5 r1
E13.5 r2
E14.5 r1
E14.5 r2
E15.5 r1
E15.5 r2
E16.5 r1
E16.5 r2
P0 r1
P0 r2
MB
E16.5 r1
E16.5 r2
P0 r1
P0 r2
HT
E16.5 r1
E16.5 r2
P0 r1
P0 r2
ST
E16.5 r1
E16.5 r2
P0 r1
P0 r2
chr1:78,092,000-78,246,000
c
Cluster
FB
MB
HB
HT
IT
ST
KD
E10.5
E11.5
E12.5
E13.5
E14.5
E15.5
E16.5
P0
AD*
AD
C1
C2
C3
C4
C5
n=121
n=42
n=121
n=42
n=58
+100kb
-100kb
mCH domain
mCH level (%)
0
0.5
d
Number of genes overlapped with mCH domains
0
50
100
150
200
250
TF
Other
mCH domain cluster
C1 19.64%
C2 3.23%
C3 16.36%
C4 53.66%
C5 39.13%
Tbx4, Notch2, Hoxb7, Hoxc9...
Erbb4, Bmp10...
Runx2, Foxf2, Hoxd3, Hoxa4...
Nr2f1, Foxp1, Ndnf, Tbr1...
Prox1, Mef2c, Lhx1, Pax3, Neurog2...
e
C1 anterior/posterior pattern specification
C2 positive regulation of cardiac muscle cell proliferation^
C3 embryonic morphogenesis
C4 regulation of neuron differentiation
C5 retinal ganglion cell axon guidance
^ adjusted p = 0.106
0
2
4
6
8
10
12
-log10(adjusted p-value)
f
Expression of genes in mCH domains relative to genes outside
TPM z-score relative to global trend
0.6
0
−0.6
FB
FB (Val.)
IT
MB
MB (Val.)
ST
HB
HB (Val.)
KD
HT
E14.5
E15.5
E16.5
P0
C1
C2
C3
C4
C5
g
C1
C2
C3
C4
C5
All
HT
HT
MB HB NT FB CF LM
MB HB FB NT
% bp in feDMRs
0
2
4
6
8
10
20
30
−100kb
mCH domain
+100kb
FB
MB
HB
NT
HT
LM
CF
KD
LG
ST
IT

Figure 4.6: **Global hypomethylation in fetal liver. a,** Example of a partially methylated domain (PMDs) in developing mouse fetal liver. The PMD location is marked by a red bar. **b,** The total bases that PMDs encompass in liver at different developmental stages. **c,** Percentage of bases in the PMDs identified in each of the liver samples (E12.5 liver, E13.5 liver etc) that are also within the PMDs identified in E15.5 liver sample. **d,** Average mCG level (mCG/CG) of PMDs and flanking regions (+/-100kb) in liver samples from different developmental stages. **e,** Histone modification profiles for H3K9me3 (top), H3K27me3 (middle) and H3K27ac (bottom) within PMDs and flanking regions (+/-100kb) in liver samples from different developmental stages. **f,** Replication timing profiling of PMDs and flanking regions (+/-100kb). The values indicate the tendency to be replicated at an earlier stage in the cell cycle. **g,** Expression of genes overlapping PMDs and flanking regions (+/-100kb) (left) compared with those with no PMD overlap (right). Two plots on the bottom show the data from a validation dataset, containing RNA-seq data generated using different protocol on the matched tissues.

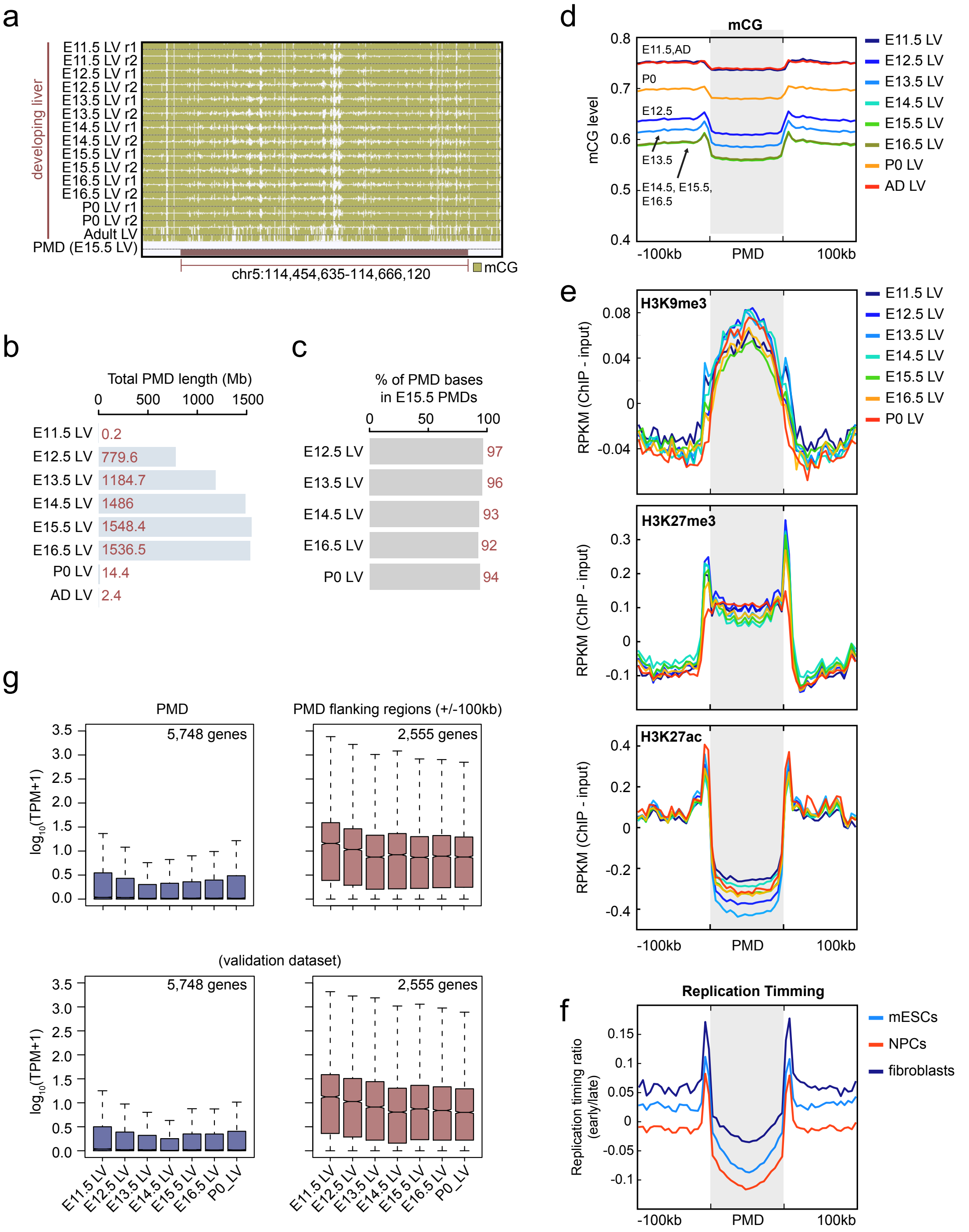
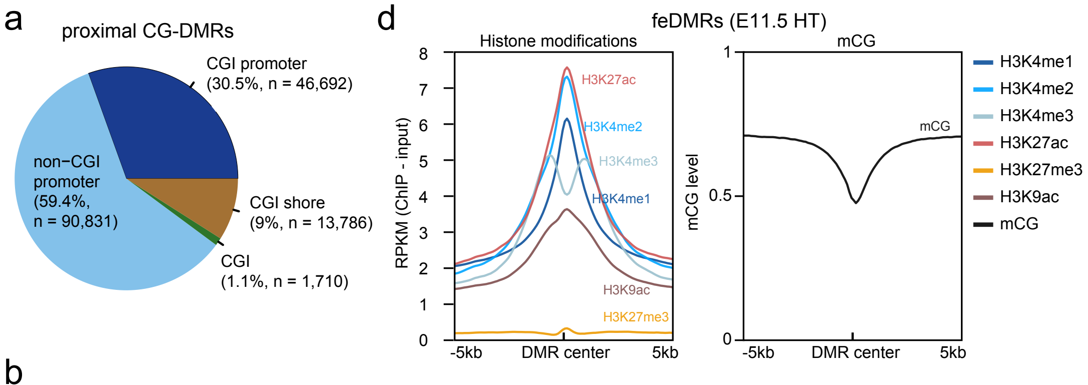

a
developing liver
E11.5 LV r1
E11.5 LV r2
E12.5 LV r1
E12.5 LV r2
E13.5 LV r1
E13.5 LV r2
E14.5 LV r1
E14.5 LV r2
E15.5 LV r1
E15.5 LV r2
E16.5 LV r1
E16.5 LV r2
P0 LV r1
P0 LV r2
Adult LV
PMD (E15.5 LV)
chr5:114,454,635-114,666,120
mCG
b
Total PMD length (Mb)
0
500
1000
1500
E11.5 LV 0.2
E12.5 LV 779.6
E13.5 LV 1184.7
E14.5 LV 1486
E15.5 LV 1548.4
E16.5 LV 1536.5
P0 LV 14.4
AD LV 2.4
c
% of PMD bases in E15.5 PMDs
0
50
100
E12.5 LV 97
E13.5 LV 96
E14.5 LV 93
E16.5 LV 92
P0 LV 94
d
mCG
mCG level
0.8
0.7
0.6
0.5
0.4
E11.5,AD
P0
E12.5
E13.5
E14.5, E15.5, E16.5
-100kb
PMD
100kb
E11.5 LV
E12.5 LV
E13.5 LV
E14.5 LV
E15.5 LV
E16.5 LV
P0 LV
AD LV
e
H3K9me3
RPKM (ChIP - input)
0.08
0.04
0
-0.04
E11.5 LV
E12.5 LV
E13.5 LV
E14.5 LV
E15.5 LV
E16.5 LV
P0 LV
H3K27me3
RPKM (ChIP - input)
0.3
0.2
0.1
0
-0.1
H3K27ac
RPKM (ChIP - input)
0.4
0.2
0
-0.2
-0.4
-100kb
PMD
100kb
f
Replication Timming
Replication timing ratio (early/late)
0.15
0.1
0.05
0
-0.05
-0.1
mESCs
NPCs
fibroblasts
-100kb
PMD
100kb
g
PMD
PMD flanking regions (+/-100kb)
5,748 genes
2,555 genes
log10(TPM+1)
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
(validation dataset)
5,748 genes
2,555 genes
E11.5 LV
E12.5 LV
E13.5 LV
E14.5 LV
E15.5 LV
E16.5 LV
P0_LV

Figure 4.7: **Categorization of CG-DMRs.** **a,** Genomic distribution of proximal CG-DMRs. **b,** Evolutionary conservation of proximal CG-DMRs overlapping with: CG islands (CGI), CGI shores, CGI promoters and non-CGI promoters. phyloP score was used to measure the degree of conservation. **c,** Chromatin signatures of fetal enhancer-linked CG-DMRs in E11.5 heart. The aggregate line plots show the average histone modification and mCG profiles of +/- 5kb regions centered at CG-DMR centers. **d,** Table summarizing the definition and number of various CG-DMR categories. Note that categories in the table are mutually exclusive.



b

CG-DMR category	Definition	Number	% of total CG-DMRs
proximal CG-DMRs	Within 1kb to promoters/CGIs/CGI shores	153,019	8.5%
distal CG-DMRs	1kb away from promoters/CGIs/CGI shores	1,655,791	91.5%
feDMR (fetal enhancer-linked CG-DMR)	Show enhancer-like chromatin signatures and were predicted as enhancers by REPTILE	487,367	22.9%
flanking distal feDMRs	Distal non-feDMRs that are within 1kb to distal feDMRs	221,960	12.3%
primed distal feDMRs	Distal CG-DMRs that show tissue-specific hypomethylation but do not belong to the above two categories	149,610	8.3%
unxDMRs (unexplained DMRs)	Distal CG-DMRs that do not belong to above three categories	868,994	48%
te-unxDMR (transposalble-elements-overlapping unxDMR)	unxDMRs that are overlapped with TEs	442,356	24.4%
nte-unxDMR (non-transposalble-elements-overlapping unxDMR)	unxDMRs that are not overlapped with TEs	426,638	23.6%

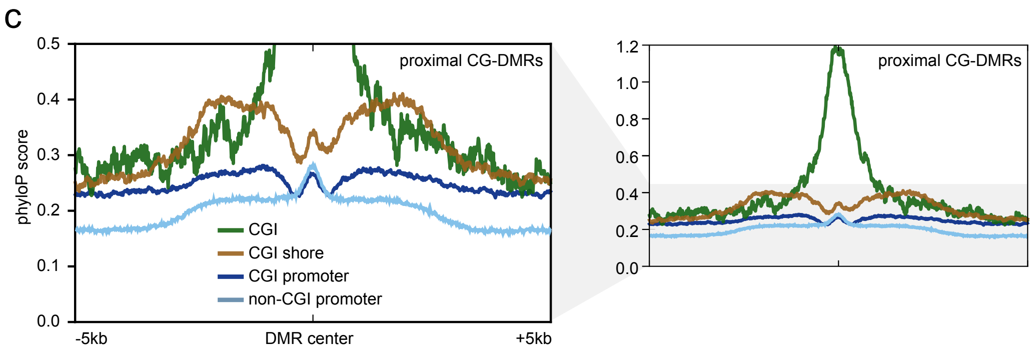


Figure 4.8: **fetal-enhancer-linked CG-DMRs (feDMRs). a,** The overlap between feDMRs identified in this study and the enhancers predicted in Yue et al[36]. Numbers in parenthesis indicate the counts of enhancers from Yue et al, whereas the remaining numbers denote the counts of feDMRs. **b,** Experimental validation results of the feDMR-overlapping elements from VISTA enhancer browser37. Different enhancer score thresholds were used for calling feDMRs for each tissue at fetal stage E11.5. Each pie shows the fraction of elements that were experimentally validated as active enhancers in matched tissue (red) or any tissue (orange) or as inactive enhancers (white) at fetal state E11.5. **c,** Enrichment of transcription factor binding motifs in feDMRs of different tissue types.

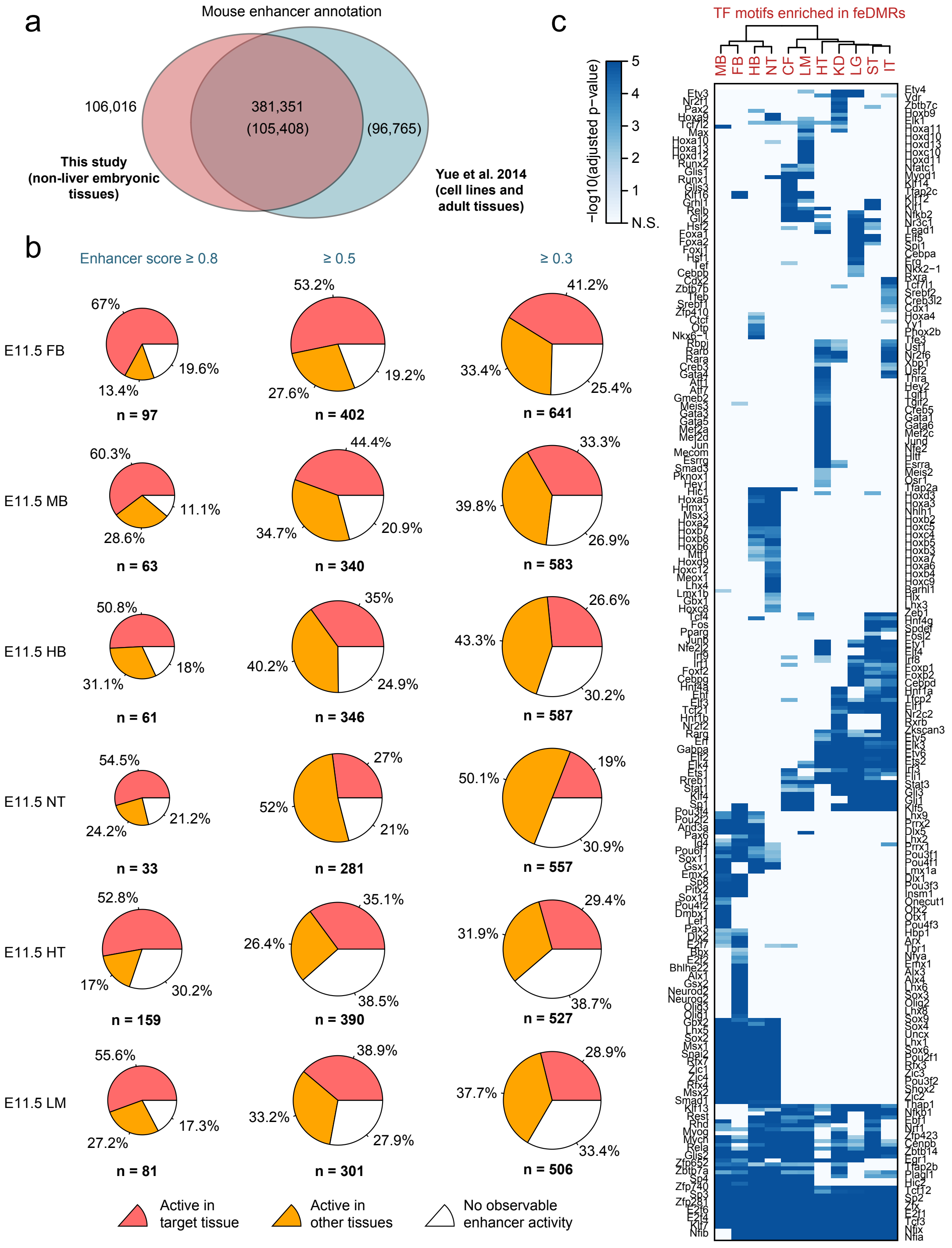

a
Mouse enhancer annotation
106,016
381,351
(105,408)
(96,765)
This study
(non-liver embryonic
tissues)
Yue et al. 2014
(cell lines and
adult tissues)
b
Enhancer score ≥ 0.8
≥ 0.5
≥ 0.3
E11.5 FB
67%
13.4%
19.6%
n = 97
53.2%
27.6%
19.2%
n = 402
41.2%
33.4%
25.4%
n = 641
E11.5 MB
60.3%
28.6%
11.1%
n = 63
44.4%
34.7%
20.9%
n = 340
33.3%
39.8%
26.9%
n = 583
E11.5 HB
50.8%
31.1%
18%
n = 61
35%
40.2%
24.9%
n = 346
26.6%
43.3%
30.2%
n = 587
E11.5 NT
54.5%
24.2%
21.2%
n = 33
27%
52%
21%
n = 281
19%
50.1%
30.9%
n = 557
E11.5 HT
52.8%
17%
30.2%
n = 159
35.1%
26.4%
38.5%
n = 390
29.4%
31.9%
38.7%
n = 527
E11.5 LM
55.6%
27.2%
17.3%
n = 81
38.9%
33.2%
27.9%
n = 301
28.9%
37.7%
33.4%
n = 506
Active in target tissue
Active in other tissues
No observable enhancer activity
c
−log10(adjusted p−value)
5
4
3
2
1
N.S.
TF motifs enriched in feDMRs
MB
FB
HB
NT
CF
LM
HT
KD
LG
ST
IT
Etv3
Nr2f1
Pax2
Hoxa9
Tcf7l2
Max
Hoxa10
Hoxa13
Hoxd12
Runx2
Glis1
Runx1
Glis3
Klf16
Grhl1
Relb
Gli2
Hsf2
Foxa1
Foxa2
Foxj1
Hsf1
Tef
Cebpb
Cdx2
Zbtb7b
Tfeb
Srebf1
Zfp410
Ctcf
Otp
Nkx6−1
Rbpj
Rarb
Rara
Creb3
Gata4
Atf1
Atf7
Gmeb2
Meis3
Gata3
Gata5
Mef2a
Mef2d
Jun
Mecom
Esrrg
Smad3
Pknox1
Hey1
Hic1
Hoxa5
Hmx1
Msx3
Hoxa2
Hoxb7
Hoxb8
Hoxb6
Mtf1
Hoxd9
Hoxc12
Meox1
Lhx4
Lmx1b
Gbx1
Hoxc8
Tcf4
Fos
Pparg
Junb
Nfe2l2
Irf9
Irf1
Foxf2
Cebpg
Hnf4a
Ehf
Elf3
Tcf21
Hnf1b
Nr2f2
Rarg
Erf
Gabpa
Elf2
Elk4
Ets1
Rreb1
Stat1
Klf4
Sp1
Pou3f4
Pou2f2
Arid3a
Pax6
Id4
Pou6f1
Sox11
Gsx1
Emx2
Sp8
Pitx2
Sox14
Pou4f2
Dmbx1
Lef1
Pax3
Dlx2
E2f7
Bbx
E2f2
Bhlhe22
Alx1
Gsx2
Neurod2
Neurog2
Olig3
Olig1
Gbx2
Lhx5
Sox2
Msx1
Snai2
Rfx7
Zic1
Zic4
Rfx4
Msx2
Smad1
Klf13
Rest
Rhd
Myog
Mycn
Rela
Glis2
Zfp652
Zbtb7a
Sp4
Zfp740
Sp3
Zfp281
E2f6
E2f4
Klf7
Nfib
Etv4
Vdr
Zbtb7c
Hoxb9
Elk1
Hoxa11
Hoxd10
Hoxd13
Hoxc10
Hoxd11
Nfatc1
Myod1
Klf14
Tfap2c
Klf12
Klf1
Nfkb2
Nr3c1
Tead1
Elf5
Spi1
Cebpa
Erg
Nkx2−1
Rxra
Tcf7l1
Srebf2
Creb3l2
Cdx1
Hoxa4
Yy1
Phox2b
Tfe3
Usf1
Nr2f6
Xbp1
Usf2
Thra
Hey2
Tgif1
Tgif2
Creb5
Gata1
Gata6
Mef2c
Jund
Nfe2
Hltf
Esrra
Meis2
Osr1
Tfap2a
Hoxd3
Hoxa3
Nhlh1
Hoxb2
Hoxc5
Hoxc4
Hoxb5
Hoxb3
Hoxa7
Hoxa6
Hoxb4
Hoxc9
Barhl1
Hlx
Lhx3
Zeb1
Hnf4g
Spdef
Fosl2
Etv1
Elf4
Irf8
Foxp1
Foxp2
Cebpd
Hnf1a
Tfcp2
Elf1
Nr2c2
Rxrb
Zkscan3
Etv5
Elk3
Etv6
Ets2
Irf3
Fli1
Stat3
Gli3
Gli1
Klf5
Lhx9
Prrx2
Dlx5
Lhx2
Prrx1
Pou3f1
Pou4f1
Lmx1a
Dlx1
Pou3f3
Insm1
Onecut1
Otx2
Otx1
Pou4f3
Hbp1
Arx
Tbr1
Nfya
Emx1
Alx3
Alx4
Lhx6
Sox3
Olig2
Lhx8
Sox9
Sox4
Uncx
Lhx1
Sox6
Pou2f1
Rfx3
Zic3
Pou3f2
Shox2
Zic2
Thap1
Nfkb1
Ebf1
Nrf1
Zfp423
Cenpb
Zbtb14
Egr1
Tfap2b
Plagl1
Hic2
Tcf12
Sp2
Zfx
E2f1
Tcf3
Nfix
Nfia

Figure 4.9: **Characterization of primed distal feDMRs and unxDMRs. a,** CG methylation (mCG) level of all primed distal fetal enhancer-linked CG-DMRs (feDMRs) in all non-liver tissues. Each row in the heatmap is one tissue sample and each column corresponds to one primed distal feDMR. Both rows and columns were clustered using hierarchical clustering. Colors bars indicate the tissue types and developmental stages of samples, respectively. **b,** mCG (left) and histone modification (right) signatures of primed distal feDMRs (blue) and feDMRs (red). Boxplots show the median and quantiles of the values in all non-liver tissues. **c,** Number of enriched transcription factor binding motifs only in feDMRs (red), only in primed distal feDMRs (orange), both (dark red) and none (grey). Only the motifs linked to expressed transcription factors (transcripts per million, TPM ¿= 10) were included. **d-e,** Similar to (a), heatmaps showing the mCG levels of unexplained CG-DMRs, including te-unxDMRs (d) and nte-unxDMRs (e).

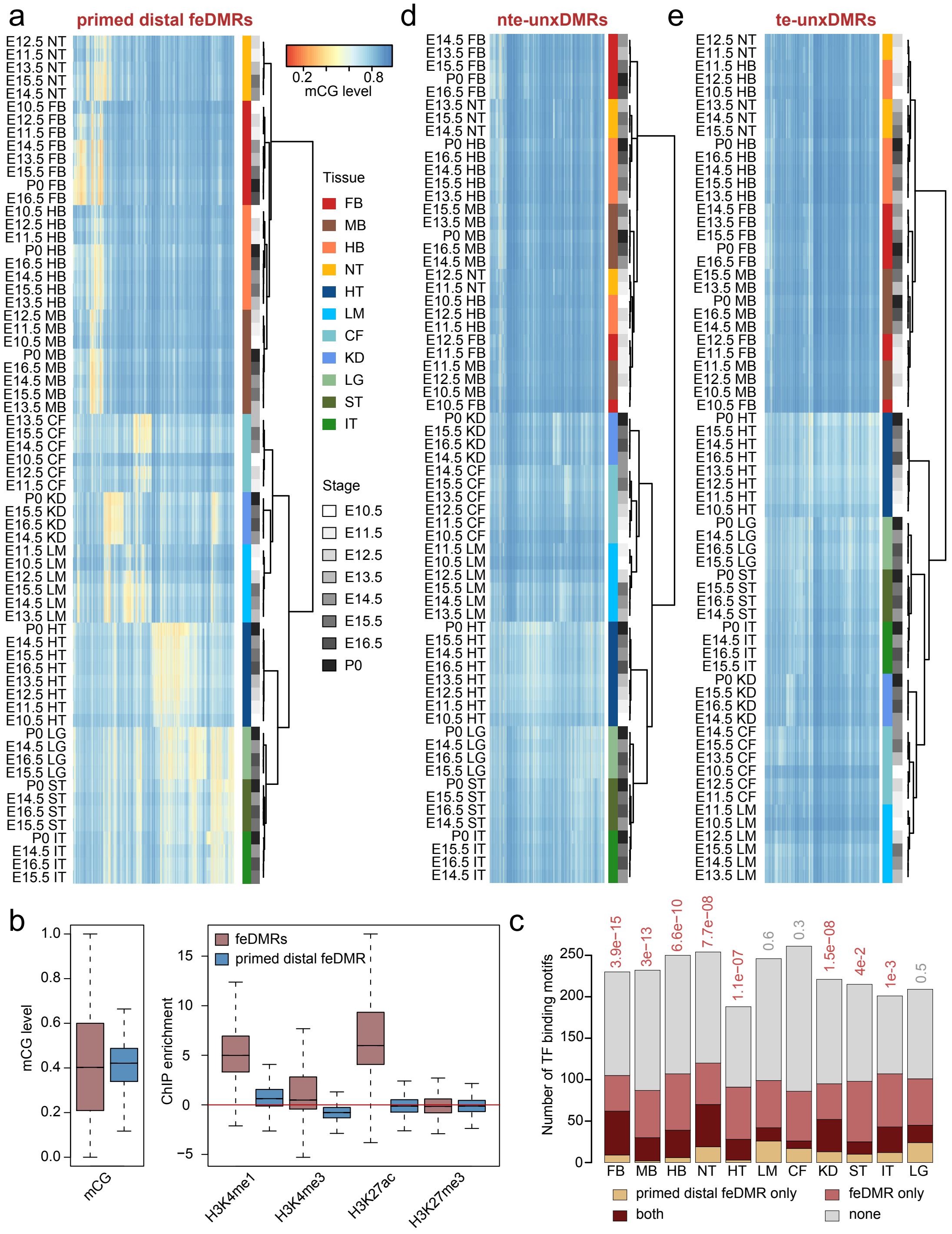

a
primed distal feDMRs
E12.5 NT
E11.5 NT
E13.5 NT
E15.5 NT
E14.5 NT
E10.5 FB
E12.5 FB
E11.5 FB
E14.5 FB
E13.5 FB
E15.5 FB
P0 FB
E16.5 FB
E10.5 HB
E12.5 HB
E11.5 HB
P0 HB
E16.5 HB
E14.5 HB
E15.5 HB
E13.5 HB
E12.5 MB
E11.5 MB
E10.5 MB
P0 MB
E16.5 MB
E14.5 MB
E15.5 MB
E13.5 MB
E13.5 CF
E15.5 CF
E14.5 CF
E10.5 CF
E12.5 CF
E11.5 CF
P0 KD
E15.5 KD
E16.5 KD
E14.5 KD
E11.5 LM
E10.5 LM
E12.5 LM
E15.5 LM
E14.5 LM
E13.5 LM
P0 HT
E14.5 HT
E15.5 HT
E16.5 HT
E13.5 HT
E12.5 HT
E11.5 HT
E10.5 HT
P0 LG
E14.5 LG
E16.5 LG
E15.5 LG
P0 ST
E14.5 ST
E16.5 ST
E15.5 ST
P0 IT
E14.5 IT
E16.5 IT
E15.5 IT
0.2
0.8
mCG level
Tissue
FB
MB
HB
NT
HT
LM
CF
KD
LG
ST
IT
Stage
E10.5
E11.5
E12.5
E13.5
E14.5
E15.5
E16.5
P0
d
nte-unxDMRs
E14.5 FB
E13.5 FB
E15.5 FB
P0 FB
E16.5 FB
E13.5 NT
E15.5 NT
E14.5 NT
P0 HB
E16.5 HB
E14.5 HB
E15.5 HB
E13.5 HB
E15.5 MB
E13.5 MB
P0 MB
E16.5 MB
E14.5 MB
E12.5 NT
E11.5 NT
E10.5 HB
E12.5 HB
E11.5 HB
E12.5 FB
E11.5 FB
E11.5 MB
E12.5 MB
E10.5 MB
E10.5 FB
P0 KD
E15.5 KD
E16.5 KD
E14.5 KD
E14.5 CF
E15.5 CF
E13.5 CF
E12.5 CF
E11.5 CF
E10.5 CF
E11.5 LM
E10.5 LM
E12.5 LM
E15.5 LM
E14.5 LM
E13.5 LM
P0 HT
E15.5 HT
E14.5 HT
E16.5 HT
E13.5 HT
E12.5 HT
E11.5 HT
E10.5 HT
P0 LG
E14.5 LG
E16.5 LG
E15.5 LG
P0 ST
E15.5 ST
E16.5 ST
E14.5 ST
P0 IT
E15.5 IT
E16.5 IT
E14.5 IT
e
te-unxDMRs
E12.5 NT
E11.5 NT
E11.5 HB
E12.5 HB
E10.5 HB
E13.5 NT
E14.5 NT
E15.5 NT
P0 HB
E16.5 HB
E14.5 HB
E15.5 HB
E13.5 HB
E14.5 FB
E13.5 FB
E15.5 FB
P0 FB
E16.5 FB
E15.5 MB
E13.5 MB
P0 MB
E16.5 MB
E14.5 MB
E12.5 FB
E11.5 FB
E11.5 MB
E12.5 MB
E10.5 MB
E10.5 FB
P0 HT
E15.5 HT
E14.5 HT
E16.5 HT
E13.5 HT
E12.5 HT
E11.5 HT
E10.5 HT
P0 LG
E14.5 LG
E16.5 LG
E15.5 LG
P0 ST
E15.5 ST
E16.5 ST
E14.5 ST
P0 IT
E14.5 IT
E16.5 IT
E15.5 IT
P0 KD
E15.5 KD
E16.5 KD
E14.5 KD
E14.5 CF
E15.5 CF
E13.5 CF
E10.5 CF
E12.5 CF
E11.5 CF
E11.5 LM
E10.5 LM
E12.5 LM
E15.5 LM
E14.5 LM
E13.5 LM
b
mCG level
1.0
0.8
0.6
0.4
0.2
0.0
mCG
ChIP enrichment
15
10
5
0
−5
feDMRs
primed distal feDMR
H3K4me1
H3K4me3
H3K27ac
H3K27me3
c
Number of TF binding motifs
250
200
150
100
50
0
3.9e−15
3e−13
6.6e−10
7.7e−08
1.1e−07
0.6
0.3
1.5e−08
4e-2
1e-3
0.5
FB
MB
HB
NT
HT
LM
CF
KD
ST
IT
LG
primed distal feDMR only
feDMR only
both
none

Figure 4.10: **mCG dynamics of tissue-specific CG-DMRs. a,** Fraction of tissue-specific CG-DMRs showing loss-of-mCG (blue) or gain-of-mCG (red) for each fetal stage. Loss-of-mCG (gain-of-mCG) event is defined as an increase (decrease) of mCG of at least 0.1. **b,** Composition of tissue-specific CG-DMRs. **c,** Percentage of loss-of-mCG events for each fetal stage. **d,** Percentage of gain-of-mCG events for each fetal stage.

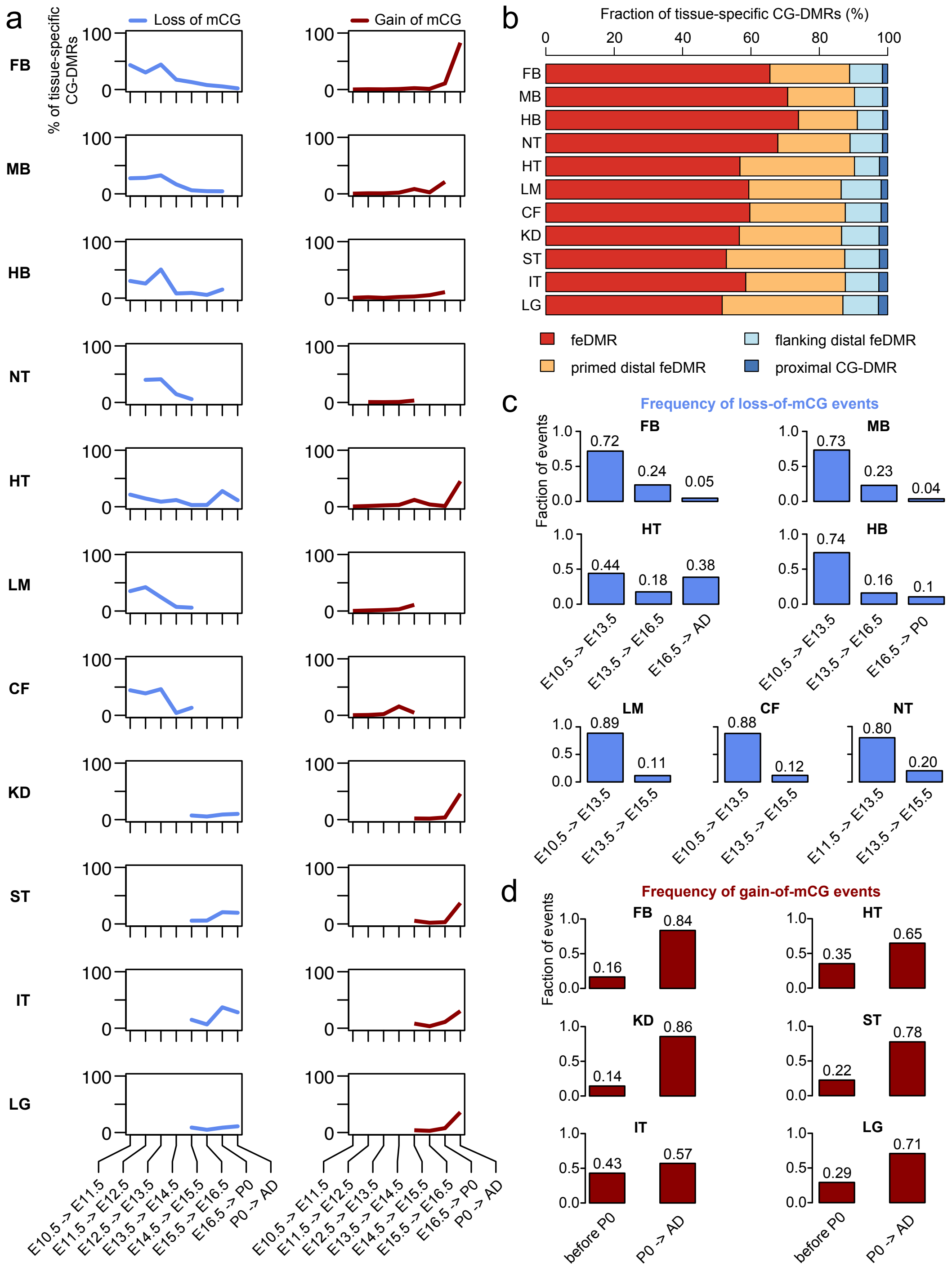

a
% of tissue-specific CG-DMRs
Loss of mCG
Gain of mCG
FB
MB
HB
NT
HT
LM
CF
KD
ST
IT
LG
100
0
E10.5 -> E11.5
E11.5 -> E12.5
E12.5 -> E13.5
E13.5 -> E14.5
E14.5 -> E15.5
E15.5 -> E16.5
E16.5 -> P0
P0 -> AD
b
Fraction of tissue-specific CG-DMRs (%)
0
20
40
60
80
100
FB
MB
HB
NT
HT
LM
CF
KD
ST
IT
LG
feDMR
primed distal feDMR
flanking distal feDMR
proximal CG-DMR
c
Frequency of loss-of-mCG events
Faction of events
FB
0.72
0.24
0.05
MB
0.73
0.23
0.04
HT
0.44
0.18
0.38
HB
0.74
0.16
0.1
E10.5 -> E13.5
E13.5 -> E16.5
E16.5 -> AD
E16.5 -> P0
LM
0.89
0.11
CF
0.88
0.12
NT
0.80
0.20
E10.5 -> E13.5
E13.5 -> E15.5
E11.5 -> E13.5
1.0
0.5
0.0
d
Frequency of gain-of-mCG events
Faction of events
FB
0.16
0.84
HT
0.35
0.65
KD
0.14
0.86
ST
0.22
0.78
IT
0.43
0.57
LG
0.29
0.71
before P0
P0 -> AD

Figure 4.11: **Link between methylation dynamics and histone modifications at tissue-specific CG-DMRs.** **a,** Fraction of tissue-specific CG-DMRs that are heavily CG methylated (mCG level ¿ 0.6). **b,** RNA abundance of genes involved in DNA methylation pathways, measured by transcripts per million (TPM). **c,** Normalized H3K27ac signals in different clusters. **d,** Top enriched ontology terms from GREAT[66] analysis for forebrain-specific CG-DMRs in different clusters. **e,** Dynamic mCG level of forebrain-specific CG-DMRs. Grey lines show the mean methylation levels of CG-DMRs in different clusters. Blue line is the mean of all clusters (grey lines).

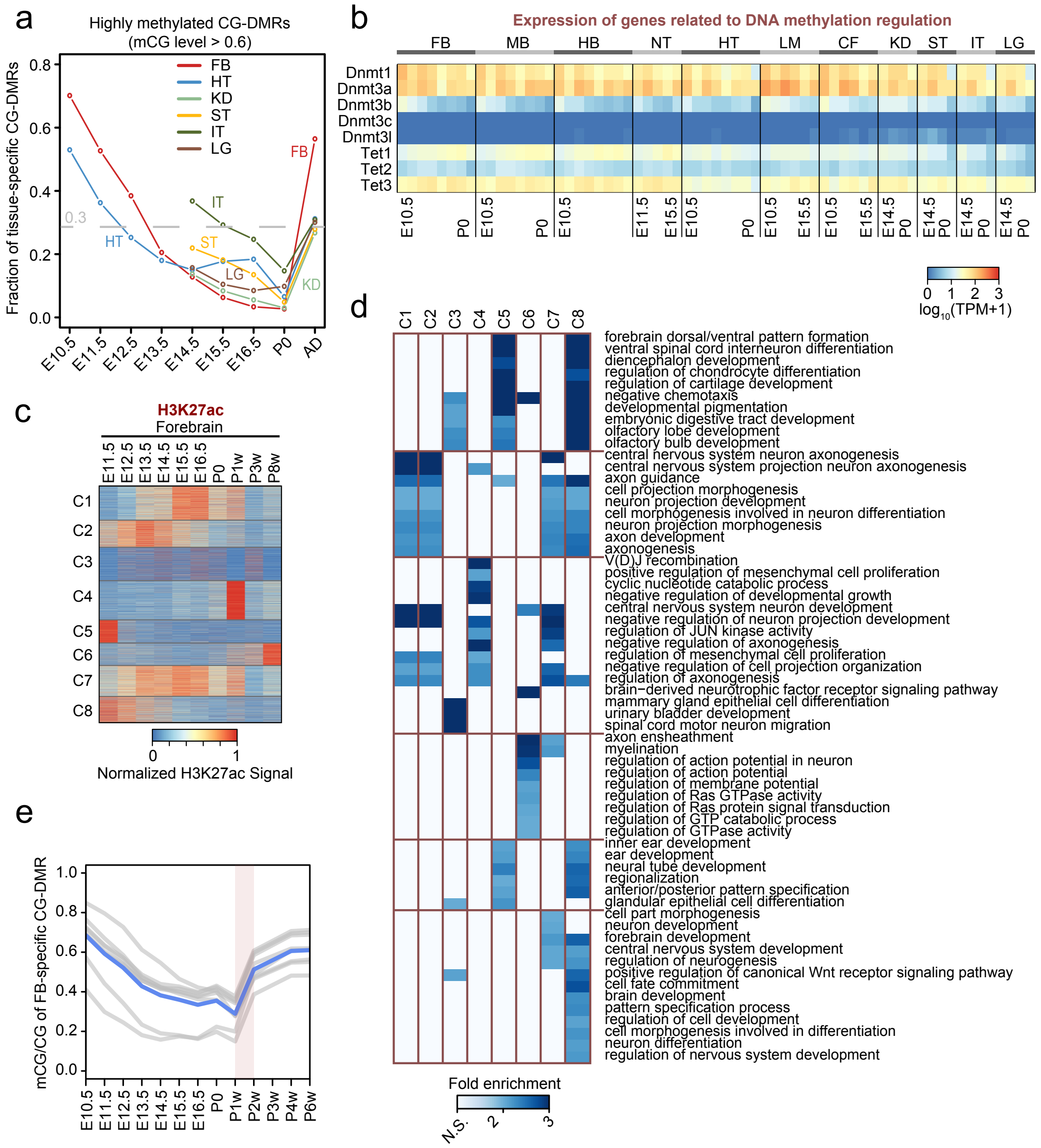

a
Highly methylated CG-DMRs
(mCG level > 0.6)
Fraction of tissue-specific CG-DMRs
0.8
0.6
0.4
0.2
0.0
0.3
FB
HT
KD
ST
IT
LG
E10.5
E11.5
E12.5
E13.5
E14.5
E15.5
E16.5
P0
AD
b
Expression of genes related to DNA methylation regulation
FB
MB
HB
NT
HT
LM
CF
KD
ST
IT
LG
Dnmt1
Dnmt3a
Dnmt3b
Dnmt3c
Dnmt3l
Tet1
Tet2
Tet3
E10.5
P0
E11.5
E15.5
E14.5
0
1
2
3
$\log_{10}$(TPM+1)
c
H3K27ac
Forebrain
E11.5
E12.5
E13.5
E14.5
E15.5
E16.5
P0
P1w
P3w
P8w
C1
C2
C3
C4
C5
C6
C7
C8
0
1
Normalized H3K27ac Signal
d
C1
C2
C3
C4
C5
C6
C7
C8
forebrain dorsal/ventral pattern formation
ventral spinal cord interneuron differentiation
diencephalon development
regulation of chondrocyte differentiation
regulation of cartilage development
negative chemotaxis
developmental pigmentation
embryonic digestive tract development
olfactory lobe development
olfactory bulb development
central nervous system neuron axonogenesis
central nervous system projection neuron axonogenesis
axon guidance
cell projection morphogenesis
neuron projection development
cell morphogenesis involved in neuron differentiation
neuron projection morphogenesis
axon development
axonogenesis
V(D)J recombination
positive regulation of mesenchymal cell proliferation
cyclic nucleotide catabolic process
negative regulation of developmental growth
central nervous system neuron development
negative regulation of neuron projection development
regulation of JUN kinase activity
negative regulation of axonogenesis
regulation of mesenchymal cell proliferation
negative regulation of cell projection organization
regulation of axonogenesis
brain−derived neurotrophic factor receptor signaling pathway
mammary gland epithelial cell differentiation
urinary bladder development
spinal cord motor neuron migration
axon ensheathment
myelination
regulation of action potential in neuron
regulation of action potential
regulation of membrane potential
regulation of Ras GTPase activity
regulation of Ras protein signal transduction
regulation of GTP catabolic process
regulation of GTPase activity
inner ear development
ear development
neural tube development
regionalization
anterior/posterior pattern specification
glandular epithelial cell differentiation
cell part morphogenesis
neuron development
forebrain development
central nervous system development
regulation of neurogenesis
positive regulation of canonical Wnt receptor signaling pathway
cell fate commitment
brain development
pattern specification process
regulation of cell development
cell morphogenesis involved in differentiation
neuron differentiation
regulation of nervous system development
Fold enrichment
N.S.
2
3
e
mCG/CG of FB-specific CG-DMR
1.0
0.8
0.6
0.4
0.2
0.0
E10.5
E11.5
E12.5
E13.5
E14.5
E15.5
E16.5
P0
P1w
P2w
P3w
P4w
P6w

Figure 4.12: **WGCNA identification of co-expression modules. a,** The scale free topology model fit (R2) (top) and the mean connectivity of the coexpression network (bottom) given different soft-thresholding powers. These two plots show how thresholds were chosen for weighted gene co-expression network analysis (WGCNA). Blue horizontal line indicates the model fit cutoff (R2 = 0.8). A soft threshold = 5 was chosen to construct the co-expression network because it is first threshold value where the model fit is greater than 0.8. **b-c,** Expression of genes in CEM12 (b) and CEM32 (c). Each row is a gene in certain module and the transcripts per million (TPM) z-scores were calculated along each row. **d,** Top enriched ontology terms of genes in co-expression modules.

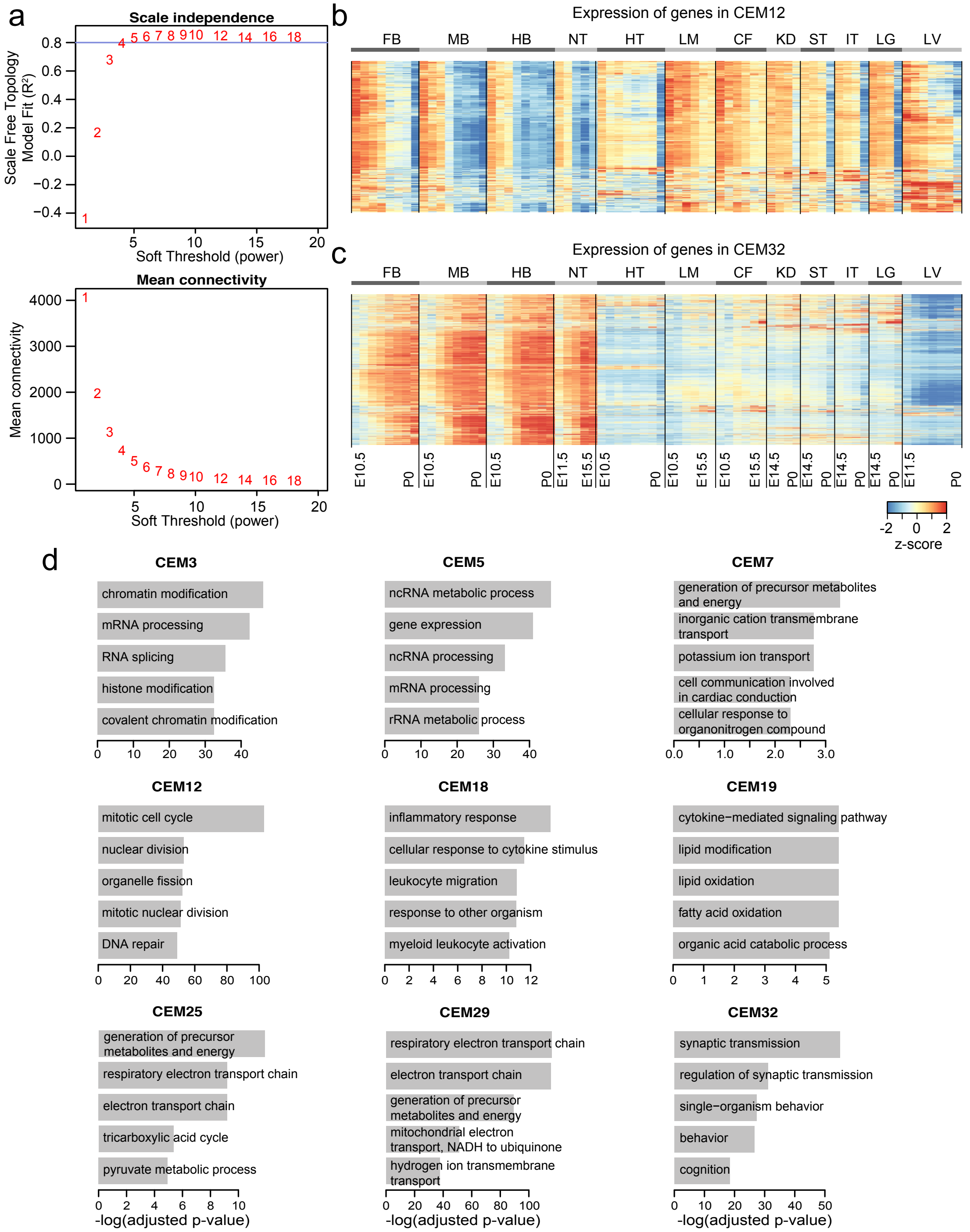
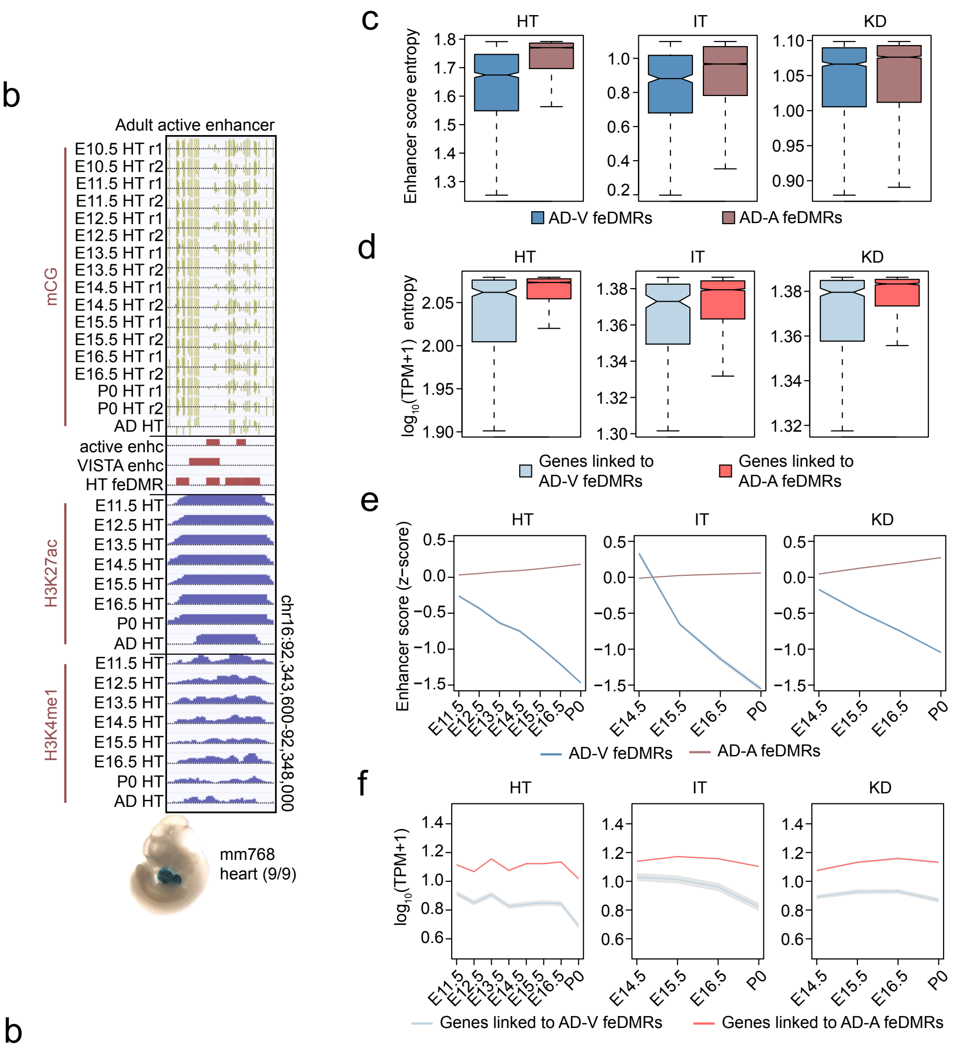

a
Scale independence
Scale Free Topology Model Fit (R²)
Soft Threshold (power)
Mean connectivity
Mean connectivity
Soft Threshold (power)
b
Expression of genes in CEM12
c
Expression of genes in CEM32
FB MB HB NT HT LM CF KD ST IT LG LV
E10.5 P0 E10.5 P0 E10.5 P0 E11.5 E15.5 E10.5 P0 E10.5 E15.5 E10.5 E15.5 E14.5 P0 E14.5 P0 E14.5 P0 E14.5 P0 E11.5 P0
-2 0 2
z-score
d
CEM3
chromatin modification
mRNA processing
RNA splicing
histone modification
covalent chromatin modification
CEM5
ncRNA metabolic process
gene expression
ncRNA processing
mRNA processing
rRNA metabolic process
CEM7
generation of precursor metabolites and energy
inorganic cation transmembrane transport
potassium ion transport
cell communication involved in cardiac conduction
cellular response to organonitrogen compound
CEM12
mitotic cell cycle
nuclear division
organelle fission
mitotic nuclear division
DNA repair
CEM18
inflammatory response
cellular response to cytokine stimulus
leukocyte migration
response to other organism
myeloid leukocyte activation
CEM19
cytokine−mediated signaling pathway
lipid modification
lipid oxidation
fatty acid oxidation
organic acid catabolic process
CEM25
generation of precursor metabolites and energy
respiratory electron transport chain
electron transport chain
tricarboxylic acid cycle
pyruvate metabolic process
CEM29
respiratory electron transport chain
electron transport chain
generation of precursor metabolites and energy
mitochondrial electron transport, NADH to ubiquinone
hydrogen ion transmembrane transport
CEM32
synaptic transmission
regulation of synaptic transmission
single−organism behavior
behavior
cognition
-log(adjusted p-value)
-log(adjusted p-value)
-log(adjusted p-value)

Figure 4.13: **Characterization of adult vestigial enhancers. a,** VISTA enhancer browser37 example of adult active enhancer (AD-A enhancer), overlapping an experimentally validated heart enhancers in the E11.5 embryo. **b,** Enhancer score entropy for feDMRs that overlap with AD-V enhancer (blue) and AD-A enhancer (red). Entropy was calculated for each CG-DMR across development. **c,** Entropy of expression (log10(TPM+1)) of genes that are nearby AD-V (blue) or AD-A feDMRs (red). The neighboring genes were linked to feDMRs. **d,** Enhancer score of AD-V (red) and AD-A feDMRs (blue). The solid lines show the average z-scores, which were calculated for stage across both AD-V and AD-A feDMRs. Grey areas surrounding solid lines indicate the standard error of the mean. **e,** Expression of genes linked to AD-V (red) and AD-A feDMRs (blue). The solid lines show the average TPM. Grey areas surrounding solid lines indicate the standard error of the mean. **f,** Fraction of AD-V CG-DMRs and AD-A CG-DMRs that were predicted as feDMRs.


b
Adult active enhancer
mCG
E10.5 HT r1
E10.5 HT r2
E11.5 HT r1
E11.5 HT r2
E12.5 HT r1
E12.5 HT r2
E13.5 HT r1
E13.5 HT r2
E14.5 HT r1
E14.5 HT r2
E15.5 HT r1
E15.5 HT r2
E16.5 HT r1
E16.5 HT r2
P0 HT r1
P0 HT r2
AD HT
active enhc
VISTA enhc
HT feDMR
H3K27ac
E11.5 HT
E12.5 HT
E13.5 HT
E14.5 HT
E15.5 HT
E16.5 HT
P0 HT
AD HT
H3K4me1
E11.5 HT
E12.5 HT
E13.5 HT
E14.5 HT
E15.5 HT
E16.5 HT
P0 HT
AD HT
chr16:92,343,600-92,348,000
mm768
heart (9/9)
c
Enhancer score entropy
HT
1.3 1.4 1.5 1.6 1.7 1.8
IT
0.2 0.4 0.6 0.8 1.0
KD
0.90 0.95 1.00 1.05 1.10
AD-V feDMRs
AD-A feDMRs
d
$\log_{10}$(TPM+1) entropy
HT
1.90 1.95 2.00 2.05
IT
1.30 1.32 1.34 1.36 1.38
KD
1.32 1.34 1.36 1.38
Genes linked to AD-V feDMRs
Genes linked to AD-A feDMRs
e
Enhancer score (z-score)
HT
IT
KD
−1.5 −1.0 −0.5 0.0 0.5
E11.5 E12.5 E13.5 E14.5 E15.5 E16.5 P0
E14.5 E15.5 E16.5 P0
E14.5 E15.5 E16.5 P0
AD-V feDMRs
AD-A feDMRs
f
$\log_{10}$(TPM+1)
HT
IT
KD
0.6 0.8 1.0 1.2 1.4
E11.5 E12.5 E13.5 E14.5 E15.5 E16.5 P0
E14.5 E15.5 E16.5 P0
E14.5 E15.5 E16.5 P0
Genes linked to AD-V feDMRs
Genes linked to AD-A feDMRs

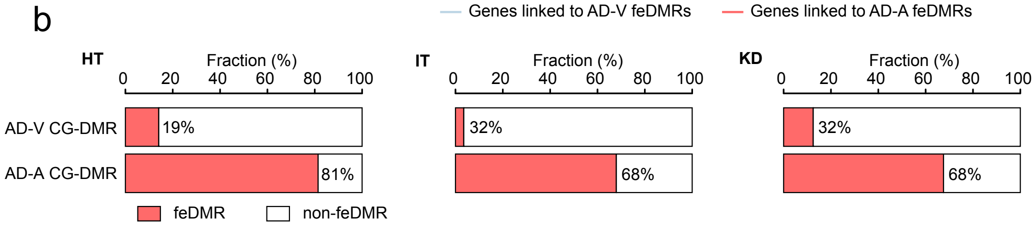

b
HT
IT
KD
Fraction (%)
0 20 40 60 80 100
AD-V CG-DMR
AD-A CG-DMR
19%
81%
32%
68%
32%
68%
feDMR
non-feDMR

Figure 4.14: **Complementary modes of gene silencing by mCG and H3K27me3 silence at adult vestigial enhancers. a-b,** Dynamic epigenetic signatures of AD-V (top) and AD-A CG-DMRs (bottom) in intestine (a) and kidney (b). For leftmost bars, red indicates CG-DMRs that were predicted as feDMR. For each CG-DMR list, the four heatmaps display (from left to right) the enhancer score, mCG level, H3K27ac signal and H3K27me3 enrichment. **c,** Barplot shows the number of AD-V non-feDMRs (top) and AD-A non-feDMRs (bottom) that are heavily CG methylated (red; mCG level ¿ 0.6), while those shown in grey are below this threshold. Numbers indicate the fractions of heavily CG methylated CG-DMRs. **d,** H3K27me3 signal at AD-V (red) and AD-A non-feDMRs (blue). Solid lines show the average TPM. Grey areas surrounding solid lines indicate the standard error of the mean. **e,** Enrichment of H3K27me3 at AD-V non-feDMRs that are heavily CG methylated (red; mCG level ¿ 0.6) and the rest (grey; mCG level ¡= 0.6).

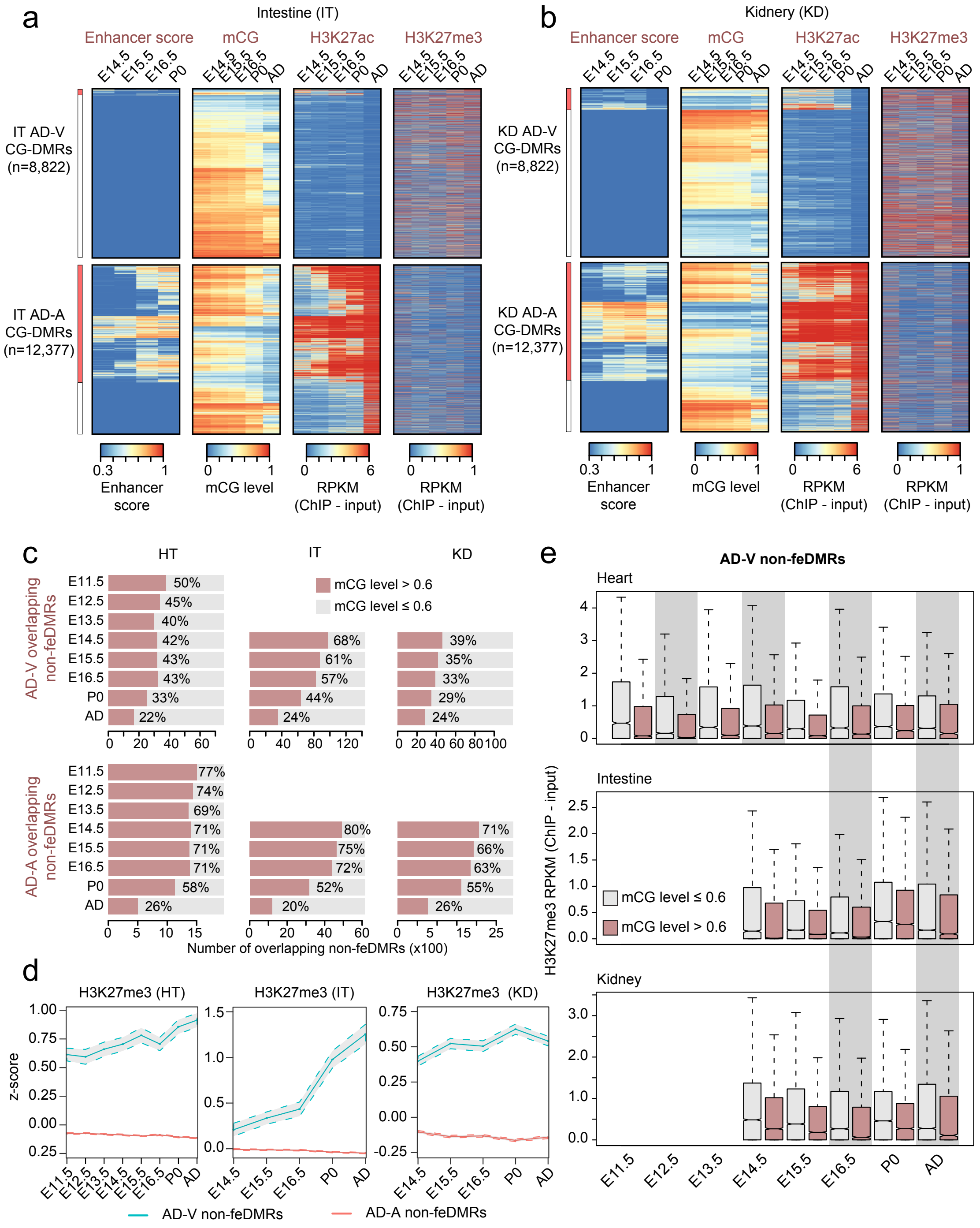
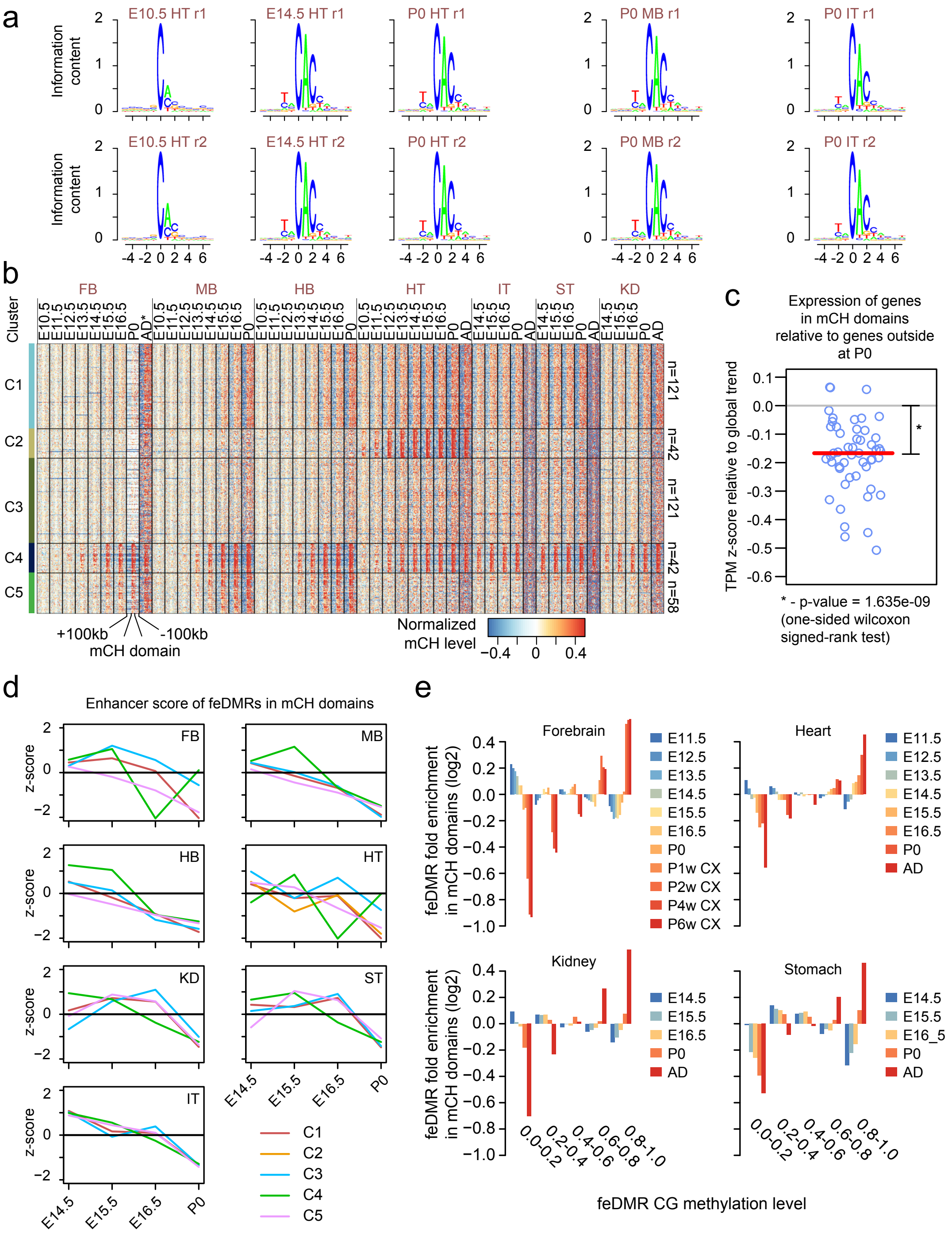

a
Intestine (IT)
Enhancer score
mCG
H3K27ac
H3K27me3
E14.5 E15.5 E16.5 P0
E14.5 E15.5 E16.5 P0 AD
IT AD-V CG-DMRs (n=8,822)
IT AD-A CG-DMRs (n=12,377)
0.3 1 Enhancer score
0 1 mCG level
0 6 RPKM (ChIP - input)
0 1 RPKM (ChIP - input)
b
Kidnery (KD)
KD AD-V CG-DMRs (n=8,822)
KD AD-A CG-DMRs (n=12,377)
c
HT
IT
KD
mCG level > 0.6
mCG level ≤ 0.6
AD-V overlapping non-feDMRs
E11.5 50%
E12.5 45%
E13.5 40%
E14.5 42% 68% 39%
E15.5 43% 61% 35%
E16.5 43% 57% 33%
P0 33% 44% 29%
AD 22% 24% 24%
0 20 40 60
0 40 80 120
0 20 40 60 80100
AD-A overlapping non-feDMRs
E11.5 77%
E12.5 74%
E13.5 69%
E14.5 71% 80% 71%
E15.5 71% 75% 66%
E16.5 71% 72% 63%
P0 58% 52% 55%
AD 26% 20% 26%
0 5 10 15
0 20 40 60
0 5 15 25
Number of overlapping non-feDMRs (x100)
d
H3K27me3 (HT)
H3K27me3 (IT)
H3K27me3 (KD)
z-score
1.00 0.75 0.50 0.25 0.00 0.25
1.5 1.0 0.5 0.0
0.75 0.50 0.25 0.00 -0.25
E11.5 E12.5 E13.5 E14.5 E15.5 E16.5 P0 AD
E14.5 E15.5 E16.5 P0 AD
AD-V non-feDMRs
AD-A non-feDMRs
e
AD-V non-feDMRs
Heart
Intestine
Kidney
H3K27me3 RPKM (ChIP - input)
mCG level ≤ 0.6
mCG level > 0.6
4 3 2 1 0
2.5 2.0 1.5 1.0 0.5 0.0
3.0 2.0 1.0 0.0
E11.5 E12.5 E13.5 E14.5 E15.5 E16.5 P0 AD

Figure 4.15: **Non-CG methylation accumulation in fetal tissues. a,** Sequence context preference for non-CG methylation (mCH). **b,** Grouping mCH domains into 5 clusters based on the dynamics of methylation accumulation. The heatmap shows normalized methylation levels of mCH domains and flanking genomic regions (up to 100kb upstream and 100kb downstream). mCH in the adult (AD) forebrain was approximated using data of frontal cortex from 6-week-old mice. **c,** Average enhancer score dynamics of feDMRs within mCH domains. Z-scores were calculated for each feDMR across development and each line shows the mean value of the mCH domains overlapping feDMRs for each cluster. **d,** Enrichment of tissue-specific feDMRs that showed different CG methylation levels in mCH domains. Colors (from blue to red) denote the stages where the mCG level was calculated. **f,** The expression of genes in mCH domains at P0 relative to the expression dynamics of genes outside mCH domains. Each circle corresponds to the value given one mCH domain cluster and one tissue. Red horizontal line indicates the median, which was tested against 0 using one-sided wilcoxon signed-rank test.


a
Information content
E10.5 HT r1
E14.5 HT r1
P0 HT r1
P0 MB r1
P0 IT r1
E10.5 HT r2
E14.5 HT r2
P0 HT r2
P0 MB r2
P0 IT r2
2
1
0
-4 -2 0 2 4 6
b
Cluster
FB
MB
HB
HT
IT
ST
KD
E10.5
E11.5
E12.5
E13.5
E14.5
E15.5
E16.5
P0
AD*
AD
C1
C2
C3
C4
C5
n=121
n=42
n=121
n=42
n=58
+100kb
-100kb
mCH domain
Normalized mCH level
−0.4
0
0.4
c
Expression of genes in mCH domains relative to genes outside at P0
TPM z-score relative to global trend
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
*
* - p-value = 1.635e-09 (one-sided wilcoxon signed-rank test)
d
Enhancer score of feDMRs in mCH domains
z-score
FB
MB
HB
HT
KD
ST
IT
2
0
−2
E14.5
E15.5
E16.5
P0
C1
C2
C3
C4
C5
e
Forebrain
Heart
Kidney
Stomach
feDMR fold enrichment in mCH domains (log2)
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
E11.5
E12.5
E13.5
E14.5
E15.5
E16.5
P0
P1w CX
P2w CX
P4w CX
P6w CX
AD
E16_5
0.0-0.2
0.2-0.4
0.4-0.6
0.6-0.8
0.8-1.0
feDMR CG methylation level

4.9 References

[1] H. Zhu, G. Wang, and J. Qian. Transcription factors as readers and effectors of DNA methylation. *Nat. Rev. Genet.*, 17(9):551–565, 08 2016.

[2] S. Domcke, A. F. Bardet, P. Adrian Ginno, D. Hartl, L. Burger, and D. Schubeler. Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature*, 528(7583):575–579, Dec 2015.

[3] J. Wan, Y. Su, Q. Song, B. Tung, O. Oyinlade, S. Liu, M. Ying, G. L. Ming, H. Song, J. Qian, H. Zhu, and S. Xia. Methylated cis-regulatory elements mediate KLF4-denpendent gene transactivation and cell migration. *Elife*, 6, May 2017.

[4] D. J. Patel and Z. Wang. Readout of epigenetic modifications. *Annu. Rev. Biochem.*, 82:81–118, 2013.

[5] D. J. Patel. A Structural Perspective on Readout of Epigenetic Histone and DNA Methylation Marks. *Cold Spring Harb Perspect Biol*, 8(3):a018754, Mar 2016.

[6] A. Bird. Dna methylation patterns and epigenetic memory. *Genes Dev*, 16:6, 2002.

[7] R. Lister, M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, J. R. Nery, L. Lee, Z. Ye, Q. M. Ngo, L. Edsall, J. Antosiewicz-Bourget, R. Stewart, V. Ruotti, A. H. Millar, J. A. Thomson, B. Ren, and J. R. Ecker. Human dna methylomes at base resolution show widespread epigenomic differences. *Nature*, 462:315–322, 2009.

[8] M. D. Schultz, Y. He, J. W. Whitaker, M. Hariharan, E. A. Mukamel, D. Leung, N. Rajagopal, J. R. Nery, M. A. Urich, H. Chen, S. Lin, Y. Lin, I. Jung, A. D. Schmitt, S. Selvaraj, B. Ren, T. J. Sejnowski, W. Wang, and J. R. Ecker. Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature*, 523(7559):212–216, Jul 2015.

[9] Ryan Lister, Eran A. Mukamel, Joseph R. Nery, Mark Urich, Clare A. Puddifoot, Nicholas D. Johnson, Jacinta Lucero, Yun Huang, Andrew J. Dwork,

Matthew D. Schultz, Miao Yu, Julian Tonti-Filippini, Holger Heyn, Shijun Hu, Joseph C. Wu, Anjana Rao, Manel Esteller, Chuan He, Fatemeh G. Haghighi, Terrence J. Sejnowski, M. Margarita Behrens, and Joseph R. Ecker. Global epigenomic reconfiguration during mammalian brain development. *Science*, Jul 2013.

[10] Michael J. Ziller, Hongcang Gu, Fabian Müller, Julie Donaghey, Linus T-Y Tsai, Oliver Kohlbacher, Philip L. De Jager, Evan D. Rosen, David A. Bennett, Bradley E. Bernstein, Andreas Gnirke, and Alexander Meissner. Charting a dynamic dna methylation landscape of the human genome. *Nature*, 500:477–481, Aug 2013.

[11] Gary C. Hon, Nisha Rajagopal, Yin Shen, David F. McCleary, Feng Yue, My D. Dang, and Bing Ren. Epigenetic memory at embryonic enhancers identified in dna methylation maps from adult mouse tissues. *Nat Genet*, 45: 1198–1206, Oct 2013.

[12] Shaohui Hu, Jun Wan, Yijing Su, Qifeng Song, Yaxue Zeng, Ha Nam Nguyen, Jaehoon Shin, Eric Cox, Hee Sool Rho, Crystal Woodard, Shuli Xia, Shuang Liu, Huibin Lyu, Guo-Li Ming, Herschel Wade, Hongjun Song, Jiang Qian, and Heng Zhu. Dna methylation presents distinct binding sites for human transcription factors. *Elife*, 2:e00726, Jan 2013.

[13] R. C. O'Malley, S. S. Huang, L. Song, M. G. Lewsey, A. Bartlett, J. R. Nery, M. Galli, A. Gallavotti, and J. R. Ecker. Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell*, 165(5):1280–1292, May 2016.

[14] Y. Yin, E. Morgunova, A. Jolma, E. Kaasinen, B. Sahu, S. Khund-Sayeed, P. K. Das, T. Kivioja, K. Dave, F. Zhong, K. R. Nitta, M. Taipale, A. Popov, P. A. Ginno, S. Domcke, J. Yan, D. Schubeler, C. Vinson, and J. Taipale. Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science*, 356(6337), May 2017.

[15] S. H. Stricker, A. Koferle, and S. Beck. From profiles to function in epigenomics. *Nat. Rev. Genet.*, 18(1):51–66, 01 2017.

[16] Wei Xie, Cathy L. Barr, Audrey Kim, Feng Yue, Ah Young Lee, James Eubanks, Emma L. Dempster, and Bing Ren. Base-resolution analyses of se-

quence and parent-of-origin dependent dna methylation in the mouse genome. *Cell*, 148:816–831, Feb 2012.

[17] Y. He and J. R. Ecker. Non-CG Methylation in the Human Genome. *Annu Rev Genomics Hum Genet*, 16:55–77, 2015.

[18] Katherine E. Varley, Jason Gertz, Kevin M. Bowling, Stephanie L. Parker, Timothy E. Reddy, Florencia Pauli-Behn, Marie K. Cross, Brian A. Williams, John A. Stamatoyannopoulos, Gregory E. Crawford, Devin M. Absher, Barbara J. Wold, and Richard M. Myers. Dynamic dna methylation across diverse human cell lines and tissues. *Genome Res*, 23:555–567, Mar 2013.

[19] J. U. Guo, Y. Su, J. H. Shin, J. Shin, H. Li, B. Xie, C. Zhong, S. Hu, T. Le, G. Fan, H. Zhu, Q. Chang, Y. Gao, G. L. Ming, and H. Song. Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. *Nat. Neurosci.*, 17(2):215–222, Feb 2014.

[20] L. Chen, K. Chen, L. A. Lavery, S. A. Baker, C. A. Shaw, W. Li, and H. Y. Zoghbi. MeCP2 binds to non-CG methylated DNA as neurons mature, influencing transcription and the timing of onset for Rett syndrome. *Proc. Natl. Acad. Sci. U.S.A.*, 112(17):5509–5514, Apr 2015.

[21] Harrison W. Gabel, Benyam Kinde, Hume Stroud, Caitlin S. Gilbert, David a. Harmin, Nathaniel R. Kastan, Martin Hemberg, Daniel H. Ebert, and Michael E. Greenberg. Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. *Nature*, 2015. ISSN 0028-0836. doi: 10.1038/nature14319. URL http://www.nature.com/nature/journal/vaop/ncurrent/full/nature14319.html?WT.ec{_}id=NATURE-20150312{#}affil-auth.

[22] S. Lagger, J. C. Connelly, G. Schweikert, S. Webb, J. Selfridge, B. H. Ramsahoye, M. Yu, C. He, G. Sanguinetti, L. C. Sowers, M. D. Walkinshaw, and A. Bird. MeCP2 recognizes cytosine methylated tri-nucleotide and di-nucleotide sequences to tune transcription in the mammalian brain. *PLoS Genet.*, 13(5):e1006793, May 2017.

[23] W. W. Tang, S. Dietmann, N. Irie, H. G. Leitch, V. I. Floros, C. R. Bradshaw, J. A. Hackett, P. F. Chinnery, and M. A. Surani. A Unique Gene Regulatory Network Resets the Human Germline Epigenome for Development. *Cell*, 161

(6):1453–1467, Jun 2015.

[24] L. Wang, J. Zhang, J. Duan, X. Gao, W. Zhu, X. Lu, L. Yang, J. Zhang, G. Li, W. Ci, W. Li, Q. Zhou, N. Aluru, F. Tang, C. He, X. Huang, and J. Liu. Programming and inheritance of parental DNA methylomes in mammals. *Cell*, 157(4):979–991, May 2014.

[25] H. Guo, P. Zhu, L. Yan, R. Li, B. Hu, Y. Lian, J. Yan, X. Ren, S. Lin, J. Li, X. Jin, X. Shi, P. Liu, X. Wang, W. Wang, Y. Wei, X. Li, F. Guo, X. Wu, X. Fan, J. Yong, L. Wen, S. X. Xie, F. Tang, and J. Qiao. The DNA methylation landscape of human early embryos. *Nature*, 511(7511):606–610, Jul 2014.

[26] Z. D. Smith, M. M. Chan, K. C. Humm, R. Karnik, S. Mekhoubad, A. Regev, K. Eggan, and A. Meissner. DNA methylation dynamics of the human preimplantation embryo. *Nature*, 511(7511):611–615, Jul 2014.

[27] Kenneth Lyons Jones. *Recognizable Patterns of Human Malformation*. Saunders, 6th edition, 2005.

[28] Diane I. Schroeder, Paul Lott, Ian Korf, and Janine M. LaSalle. Large-scale methylation domains mark a functional subset of neuronally expressed genes. *Genome Res*, 21:1583–1591, Oct 2011.

[29] Diane I. Schroeder, John D. Blair, Paul Lott, Hung On Ken Yu, Danna Hong, Florence Crary, Paul Ashwood, Cheryl Walker, Ian Korf, Wendy P. Robinson, and Janine M. LaSalle. The human placenta methylome. *Proc Natl Acad Sci U S A*, 110:6037–6042, Apr 2013.

[30] B. P. Berman, D. J. Weisenberger, J. F. Aman, T. Hinoue, Z. Ramjan, Y. Liu, H. Noushmehr, C. P. Lange, Dijk CM van, R. A. Tollenaar, Den Berg D. Van, and P. W. Laird. Regions of focal dna hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nat Genet*, 44:40–46, Jan 2012.

[31] G. C. Hon, R. D. Hawkins, O. L. Caballero, C. Lo, R. Lister, M. Pelizzola, A. Valsesia, Z. Ye, S. Kuan, L. E. Edsall, A. A. Camargo, B. J. Stevenson,

J. R. Ecker, V. Bafna, R. L. Strausberg, A. J. Simpson, and B. Ren. Global dna hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. *Genome Res*, 22:246–258, Feb 2012.

[32] Yupeng He, David U Gorkin, Diane E Dickel, Joseph R Nery, Rosa G Castanon, Ah Young Lee, Yin Shen, Axel Visel, Len A Pennacchio, Bing Ren, and Joseph R Ecker. Improved regulatory element prediction based on tissue-specific local epigenomic signatures. *Proceedings of the National Academy of Sciences*, 2017. doi: 10.1073/pnas.1618353114. URL http://www.pnas.org/content/early/2017/02/07/1618353114.abstract.

[33] Nathaniel D Heintzman, Rhona K Stuart, Gary Hon, Yutao Fu, Christina W Ching, R David Hawkins, Leah O Barrera, Sara Van Calcar, Chunxu Qu, Keith a Ching, Wei Wang, Zhiping Weng, Roland D Green, Gregory E Crawford, and Bing Ren. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics*, 39(3):311–318, 2007. ISSN 1061-4036. doi: 10.1038/ng1966.

[34] Nathaniel D Heintzman, Gary C Hon, R David Hawkins, Pouya Kheradpour, Alexander Stark, Lindsey F Harp, Zhen Ye, Leonard K Lee, Rhona K Stuart, Christina W Ching, Keith a Ching, Jessica E Antosiewicz, Hui Liu, Xinmin Zhang, Roland D Green, Ron Stewart, James a Thomson, and Gregory E Crawford. Histone modification at human enhancers reflect global cell-type specific gene expression. *Nature*, 459(7243):108–112, 2009. doi: 10.1038/nature07829.Histone.

[35] R. David Hawkins, Gary C. Hon, Leonard K. Lee, Queminh Ngo, Ryan Lister, Mattia Pelizzola, Lee E. Edsall, Samantha Kuan, Ying Luu, Sarit Klugman, Jessica Antosiewicz-Bourget, Zhen Ye, Celso Espinoza, Saurabh Agarwahl, Li Shen, Victor Ruotti, Wei Wang, Ron Stewart, James a. Thomson, Joseph R. Ecker, and Bing Ren. Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell*, 6(5):479–491, 2010. ISSN 19345909. doi: 10.1016/j.stem.2010.03.018. URL http://dx.doi.org/10.1016/j.stem.2010.03.018.

[36] F Yue, Y Cheng, A Breschi, J Vierstra, W Wu, T Ryba, R Sandstrom, Z Ma, C Davis, B D Pope, Y Shen, D D Pervouchine, S Djebali, R E Thurman, R Kaul, E Rynes, A Kirilusha, G K Marinov, B A Williams, D Trout, H Amrhein, K Fisher-Aylor, I Antoshechkin, G DeSalvo, L H See, M Fastuca,

J Drenkow, C Zaleski, A Dobin, P Prieto, J Lagarde, G Bussotti, A Tanzer, O Denas, K Li, M A Bender, M Zhang, R Byron, M T Groudine, D McCleary, L Pham, Z Ye, S Kuan, L Edsall, Y C Wu, M D Rasmussen, M S Bansal, M Kellis, C A Keller, C S Morrissey, T Mishra, D Jain, N Dogan, R S Harris, P Cayting, T Kawli, A P Boyle, G Euskirchen, A Kundaje, S Lin, Y Lin, C Jansen, V S Malladi, M S Cline, D T Erickson, V M Kirkup, K Learned, C A Sloan, K R Rosenbloom, B Lacerda de Sousa, K Beal, M Pignatelli, P Flicek, J Lian, T Kahveci, D Lee, W J Kent, M Ramalho Santos, J Herrero, C Notredame, A Johnson, S Vong, K Lee, D Bates, F Neri, M Diegel, T Canfield, P J Sabo, M S Wilken, T A Reh, E Giste, A Shafer, T Kutyavin, E Haugen, D Dunn, A P Reynolds, S Neph, R Humbert, R S Hansen, M De Bruijn, L Selleri, A Rudensky, S Josefowicz, R Samstein, E E Eichler, S H Orkin, D Levasseur, T Papayannopoulou, K H Chang, A Skoultchi, S Gosh, C Disteche, P Treuting, Y Wang, M J Weiss, G A Blobel, X Cao, S Zhong, T Wang, P J Good, R F Lowdon, L B Adams, X Q Zhou, M J Pazin, E A Feingold, B Wold, J Taylor, A Mortazavi, S M Weissman, J A Stamatoyannopoulos, M P Snyder, R Guigo, T R Gingeras, D M Gilbert, R C Hardison, M A Beer, B Ren, and Encode Consortium Mouse. A comparative encyclopedia of DNA elements in the mouse genome. *Nature*, 515(7527):355–364, 2014. ISSN 0028-0836. doi: 10.1038/nature13992. URL http://www.ncbi.nlm.nih.gov/pubmed/25409824{%}5Cnhttp://www.nature.com/nature/journal/v515/n7527/pdf/nature13992.pdf.

[37] Axel Visel, Simon Minovitsky, Inna Dubchak, and Len a. Pennacchio. VISTA Enhancer Browser - A database of tissue-specific human enhancers. *Nucleic Acids Research*, 35(SUPPL. 1):88–92, 2007. ISSN 03051048. doi: 10.1093/nar/gkl822.

[38] Luca Magnani, Jerome Eeckhoute, and Mathieu Lupien. Pioneer factors: directing transcriptional regulators within the chromatin environment. *Trends in genetics : TIG*, 27(11):465–474, nov 2011. ISSN 0168-9525 (Print). doi: 10.1016/j.tig.2011.07.002.

[39] Richard I Sherwood, Tatsunori Hashimoto, Charles W O'Donnell, Sophia Lewis, Amira A Barkal, John Peter van Hoff, Vivek Karun, Tommi Jaakkola, and David K Gifford. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat Biotech*, 32(2):171–178, feb 2014. ISSN 1087-0156. URL http://dx.doi.org/10.1038/nbt.2798http://10.0.4.14/nbt.2798http://www.nature.com/nbt/journal/v32/n2/abs/nbt.2798.html{#}supplementary-information.

[40] Alain Nepveu. Role of the multifunctional CDP/Cut/Cux homeodomain transcription factor in regulating differentiation, cell growth and development. *Gene*, 270(12):1–15, may 2001. ISSN 0378-1119. doi: https://doi.org/10.1016/S0378-1119(01)00485-1. URL http://www.sciencedirect.com/science/article/pii/S0378111901004851.

[41] Eliezer Calo and Joanna Wysocka. Modification of enhancer chromatin: what, how and why? *Molecular cell*, 49(5):10.1016/j.molcel.2013.01.038, mar 2013. ISSN 1097-2765. doi: 10.1016/j.molcel.2013.01.038. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3857148/.

[42] Edward B Chuong, Nels C Elde, and Cedric Feschotte. Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet*, 18(2):71–86, feb 2017. ISSN 1471-0056. URL http://dx.doi.org/10.1038/nrg.2016.139http://10.0.4.14/nrg.2016.139.

[43] Alisa Mo, Eran A. Mukamel, Fred P. Davis, Chongyuan Luo, Gilbert L. Henry, Serge Picard, Mark A. Urich, Joseph R. Nery, Terrence J. Sejnowski, Ryan Lister, Sean R. Eddy, Joseph R. Ecker, and Jeremy Nathans. Epigenomic Signatures of Neuronal Diversity in the Mammalian Brain. *Neuron*, 86(6):1369–1384, 2015. ISSN 10974199. doi: 10.1016/j.neuron.2015.05.018. URL http://dx.doi.org/10.1016/j.neuron.2015.05.018.

[44] Jeff Vierstra, Eric Rynes, Richard Sandstrom, Miaohua Zhang, Theresa Canfield, R Scott Hansen, Sandra Stehling-sun, Peter J Sabo, Rachel Byron, Richard Humbert, Robert E Thurman, Audra K Johnson, Shinny Vong, Kristen Lee, Daniel Bates, Fidencio Neri, Morgan Diegel, Erika Giste, Eric Haugen, Douglas Dunn, Matthew S Wilken, Steven Josefowicz, Robert Samstein, Kai-hsin Chang, Evan E Eichler, Marella De Bruijn, Thomas A Reh, Arthur Skoultchi, Alexander Rudensky, Stuart H Orkin, Thalia Papayannopoulou, Piper M Treuting, Licia Selleri, and Rajinder Kaul. Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. 346(6212):1007–1013, 2014.

[45] AlexS. Nord, MatthewJ. Blow, Catia Attanasio, JenniferA. Akiyama, Amy Holt, Roya Hosseini, Sengthavy Phouanenavong, Ingrid Plajzer-Frick, Malak Shoukry, Veena Afzal, JohnL.R. Rubenstein, EdwardM. Rubin, LenA. Pennacchio, and Axel Visel. Rapid and Pervasive Changes in Genome-wide Enhancer Usage during Mammalian Development. *Cell*, 155(7):1521–1531,

dec 2013. ISSN 0092-8674. doi: http://dx.doi.org/10.1016/j.cell.2013.11.033. URL http://www.sciencedirect.com/science/article/pii/S0092867413014840.

[46] Peter Langfelder and Steve Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1):559, 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-559. URL http://dx.doi.org/10.1186/1471-2105-9-559.

[47] Howard Cedar and Yehudit Bergman. Linking DNA methylation and histone modification: patterns and paradigms. *Nature reviews. Genetics*, 10(mAy): 295–304, 2009. ISSN 1471-0056. doi: 10.1038/nrg2540.

[48] Chongyuan Luo, Madeline A Lancaster, Rosa Castanon, Joseph R Nery, Juergen A Knoblich, and Joseph R Ecker. Cerebral Organoids Recapitulate Epigenomic Signatures of the Human Fetal Brain. *Cell Reports*, 17(12):3369–3384, dec 2016. ISSN 2211-1247. doi: http://doi.org/10.1016/j.celrep.2016.12.001. URL http://www.sciencedirect.com/science/article/pii/S2211124716316722.

[49] Yuanbiao Guo, Xuequn Zhang, Jian Huang, Yan Zeng, Wei Liu, Chao Geng, Ka Wan Li, Dong Yang, Songfeng Wu, Handong Wei, Zeguang Han, Xiaohong Qian, Ying Jiang, and Fuchu He. Relationships between Hematopoiesis and Hepatogenesis in the Midtrimester Fetal Liver Characterized by Dynamic Transcriptomic and Proteomic Profiles. *PLOS ONE*, 4(10):e7641, oct 2009. URL https://doi.org/10.1371/journal.pone.0007641.

[50] Alexander Medvinsky, Stanislav Rybtsov, and Samir Taoudi. Embryonic origin of the adult hematopoietic system: advances and questions. *Development*, 138(6):1017–1031, 2011. ISSN 0950-1991. doi: 10.1242/dev.040998. URL http://dev.biologists.org/content/138/6/1017.

[51] Mark A Urich, Joseph R Nery, Ryan Lister, Robert J Schmitz, and Joseph R Ecker. MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. *Nat. Protocols*, 10(3):475–483, mar 2015. ISSN 1754-2189. URL http://dx.doi.org/10.1038/nprot.2014.114http://10.0.4.14/nprot.2014.114.

[52] Laurence R. Meyer, Ann S. Zweig, Angie S. Hinrichs, Donna Karolchik, Robert M. Kuhn, Matthew Wong, Cricket A. Sloan, Kate R. Rosenbloom,

Greg Roe, Brooke Rhead, Brian J. Raney, Andy Pohl, Venkat S. Malladi, Chin H. Li, Brian T. Lee, Katrina Learned, Vanessa Kirkup, Fan Hsu, Steve Heitner, Rachel A. Harte, Maximilian Haeussler, Luvina Guruvadoo, Mary Goldman, Belinda M. Giardine, Pauline A. Fujita, Timothy R. Dreszer, Mark Diekhans, Melissa S. Cline, Hiram Clawson, Galt P. Barber, David Haussler, and W. James Kent. The ucsc genome browser database: extensions and updates 2013. *Nucleic Acids Res*, 41:D64–D69, Jan 2013.

[53] Hong Ma, Robert Morey, Ryan C. O'Neil, Yupeng He, Brittany Daughtry, Matthew D. Schultz, Manoj Hariharan, Joseph R. Nery, Rosa Castanon, Karen Sabatini, Rathi D. Thiagarajan, Masahito Tachibana, Eunju Kang, Rebecca Tippner-Hedges, Riffat Ahmed, Nuria Marti Gutierrez, Crystal Van Dyken, Alim Polat, Atsushi Sugawara, Michelle Sparman, Sumita Gokhale, Paula Amato, Don P Wolf, Joseph R. Ecker, Louise C. Laurent, and Shoukhrat Mitalipov. Abnormalities in human pluripotent cells due to reprogramming mechanisms. *Nature*, 511(7508):177–83, 2014. ISSN 1476-4687. doi: 10.1038/nature13551. URL http://www.nature.com/doifinder/10.1038/nature13551{%}5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/25008523.

[54] Matthew D. Schultz, Robert J. Schmitz, and Joseph R. Ecker. 'leveling' the playing field for analyses of single-base resolution dna methylomes. *Trends Genet*, 28:583–585, Dec 2012.

[55] Heng Li and Richard Durbin. Fast and accurate short read alignment with BurrowsWheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009. doi: 10.1093/bioinformatics/btp324. URL http://bioinformatics.oxfordjournals.org/content/25/14/1754.abstract.

[56] Fidel Ramírez, Devon P Ryan, Björn Grüning, Vivek Bhardwaj, Fabian Kilpert, Andreas S Richter, Steffen Heyne, Friederike Dündar, and Thomas Manke. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research*, 44(W1):W160–W165, 2016. doi: 10.1093/nar/gkw257. URL http://nar.oxfordjournals.org/content/44/W1/W160.abstract.

[57] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, and X Shirley Liu. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):1–9, 2008. ISSN 1474-760X. doi: 10.1186/gb-2008-9-9-r137.

URL http://dx.doi.org/10.1186/gb-2008-9-9-r137.

[58] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, 29(1):15–21, jan 2013. ISSN 1367-4811 (Electronic). doi: 10.1093/bioinformatics/bts635.

[59] Jennifer Harrow, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L. Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, If Barnes, Alexandra Bignell, Veronika Boychenko, Toby Hunt, Mike Kay, Gaurab Mukherjee, Jeena Rajan, Gloria Despacio-Reyes, Gary Saunders, Charles Steward, Rachel Harte, Michael Lin, Cédric Howald, Andrea Tanzer, Thomas Derrien, Jacqueline Chrast, Nathalie Walters, Suganthi Balasubramanian, Baikang Pei, Michael Tress, Jose Manuel Rodriguez, Iakes Ezkurdia, Jeltje van Baren, Michael Brent, David Haussler, Manolis Kellis, Alfonso Valencia, Alexandre Reymond, Mark Gerstein, Roderic Guigó, and Tim J. Hubbard. Gencode: the reference human genome annotation for the encode project. *Genome Res*, 22:1760–1774, Sep 2012.

[60] Bo Li and Colin N Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12 (1):323, 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-323. URL http://dx.doi.org/10.1186/1471-2105-12-323.

[61] Aaron R Quinlan and Ira M Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010. doi: 10.1093/bioinformatics/btq033. URL http://bioinformatics.oxfordjournals.org/content/26/6/841.abstract.

[62] Peter Rousseeuw. Least Median of Squares Regression. *Journal of the American Statistical Association*, 79(388):871–880, 1984. ISSN 0162-1459. doi: 10.1080/01621459.1984.10477105. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1984.10477105.

[63] Charles E Grant, Timothy L Bailey, and William Stafford Noble. FIMO: scanning for occurrences of a given motif. *Bioinformatics (Oxford, England)*, 27(7):1017–1018, apr 2011. ISSN 1367-4811 (Electronic). doi: 10.1093/bioinformatics/btr064.

[64] Danielle Welter, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, and Helen Parkinson. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 42(D1):D1001, 2013. doi: 10.1093/nar/gkt1229. URL +http://dx.doi.org/10.1093/nar/gkt1229.

[65] Katherine S Pollard, Melissa J Hubisz, Kate R Rosenbloom, and Adam Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20(1):110–121, 2010. doi: 10.1101/gr.097857.109. URL http://genome.cshlp.org/content/20/1/110.abstract.

[66] C. Y. McLean, D. Bristor, M. Hiller, S. L. Clarke, B. T. Schaar, C. B. Lowe, A. M. Wenger, and G. Bejerano. Great improves functional interpretation of cis-regulatory regions. *Nat Biotechnol*, 28:495–501, May 2010.

[67] Edward Y Chen, Christopher M Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R Clark, and Avi Ma'ayan. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC bioinformatics*, 14:128, apr 2013. ISSN 1471-2105 (Electronic). doi: 10.1186/1471-2105-14-128.

[68] Maxim V Kuleshov, Matthew R Jones, Andrew D Rouillard, Nicolas F Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L Jenkins, Kathleen M Jagodnik, Alexander Lachmann, Michael G McDermott, Caroline D Monteiro, Gregory W Gundersen, and Avi Ma'ayan. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, 44(W1):W90–7, jul 2016. ISSN 1362-4962 (Electronic). doi: 10.1093/nar/gkw377.

[69] Wei Xie, Matthew D. Schultz, Ryan Lister, Zhonggang Hou, Nisha Rajagopal, Pradipta Ray, John W. Whitaker, Shulan Tian, R. David Hawkins, Danny Leung, Hongbo Yang, Tao Wang, Ah Young Lee, Scott A. Swanson, Jiuchun Zhang, Yun Zhu, Audrey Kim, Joseph R. Nery, Mark A. Urich, Samantha Kuan, Chia-An Yen, Sarit Klugman, Pengzhi Yu, Kran Suknuntha, Nicholas E. Propson, Huaming Chen, Lee E. Edsall, Ulrich Wagner, Yan Li, Zhen Ye, Ashwinikumar Kulkarni, Zhenyu Xuan, Wen-Yu Chung, Neil C. Chi, Jessica E. Antosiewicz-Bourget, Igor Slukvin, Ron Stewart, Michael Q. Zhang, Wei Wang, James A. Thomson, Joseph R. Ecker, and Bing Ren. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell*,

153:1134–1148, May 2013.

[70] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[71] Yuhua Fu, Pin Lv, Guoquan Yan, Hui Fan, Lu Cheng, Feng Zhang, Yongjun Dang, Hao Wu, and Bo Wen. MacroH2A1 associates with nuclear lamina and maintains chromatin architecture in mouse liver cells. *Scientific Reports*, 5:17186, nov 2015. URL http://dx.doi.org/10.1038/srep17186http://10.0.4.14/srep17186http://www.nature.com/articles/srep17186{#}supplementary-information.

[72] Nodin Weddington, Alexander Stuy, Ichiro Hiratani, Tyrone Ryba, Tomoki Yokochi, and David M Gilbert. ReplicationDomain: a visualization tool and comparative database for genome-wide replication timing data. *BMC Bioinformatics*, 9(1):530, 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-530. URL http://dx.doi.org/10.1186/1471-2105-9-530.

[73] Oliver Bembom. *seqLogo: Sequence logos for DNA sequence alignments*. R package version 1.36.0.

[74] Hong-Mei Zhang, Hu Chen, Wei Liu, Hui Liu, Jing Gong, Huili Wang, and An-Yuan Guo. AnimalTFDB: a comprehensive animal transcription factor database. *Nucleic Acids Research*, 40(D1):D144, 2012. doi: 10.1093/nar/gkr965. URL +http://dx.doi.org/10.1093/nar/gkr965.

Chapter 5

Conclusion

In my thesis works, I describe the efforts to profile the DNA methylation landscape of a variety of human and mouse tissues, develop new computational tool for interpreting the DNA methylation data. I and my colleagues studied the tissue-specific DNA methylation on the context of fetal development and characterized the dynamic DNA methylation regulation. Also, I developed a computational algorithm, REPTILE, which integrates DNA methylation and histone modification data and generates accurate, high-resolution enhancer predictions. These projects led to several concrete deliverables:

1. The genome-wide, base-resolution DNA methylation profiling of human and mouse tissues, along with the data generated in other studies[1, 2, 3, 4], lies the foundation of interrogating this epigenetic modification across a variety of cell types and tissues. These comprehensive maps serve as the methylome baseline of normal tissues, fetal development etc, which can be valuable for studying DNA methylation changes related to diseases. Specifically, as shown

in Chapter 4, the methylomes of developing mouse fetal tissues will be useful for studying human birth defects.

2. In-depth analysis of the human and mouse methylomes reveals principles of DNA methylation regulation in different tissues and at different development stages. Similar to the results from Ziller et al[1] and Hon et al[5], we found that the genomic regions showing tissue-specific DNA methylation are strongly enriched for regulatory elements, especially enhancers. During fetal development stages, these enhancer regions undergo major demethylation whereas the trend is reversed after birth.

3. The results of REPTILE algorithm demonstrate that DNA methylation can be combined with histone modifications to generate accurate enhancer predictions. We expected this tool to be useful for interpreting the current epigenomic datasets and generating high-resolution enhancer annotations for a variety of tissues and cell types.

4. We applied REPTILE to identify enhancers for developing mouse fetal tissues. Some of the enhancers can be validated by *in vivo* experiments. Furthermore, the human orthologs of these enhancers are enriched for genetic risk factors associated with human diseases. Such spatiotemporal enhancer annotation of mouse embryo will be useful for studying mammalian development and also birth defects.

5. Lastly, our studies reveal that the previously understudied non-CG methylation are present in human and mouse tissues that were not known to contain non-CG methylation[6]. Non-CG methylation accumulates tissue-specifically

in the bodies of genes that encode transcription factors and is associated with the repression of these genes, which may be related to the binding of MeCP2 on non-CG methylated DNA[7, 8, 9, 10].

5.1 References

[1] Michael J. Ziller, Hongcang Gu, Fabian Müller, Julie Donaghey, Linus T-Y Tsai, Oliver Kohlbacher, Philip L. De Jager, Evan D. Rosen, David A. Bennett, Bradley E. Bernstein, Andreas Gnirke, and Alexander Meissner. Charting a dynamic dna methylation landscape of the human genome. *Nature*, 500:477–481, Aug 2013.

[2] Ryan Lister, Eran A. Mukamel, Joseph R. Nery, Mark Urich, Clare A. Puddifoot, Nicholas D. Johnson, Jacinta Lucero, Yun Huang, Andrew J. Dwork, Matthew D. Schultz, Miao Yu, Julian Tonti-Filippini, Holger Heyn, Shijun Hu, Joseph C. Wu, Anjana Rao, Manel Esteller, Chuan He, Fatemeh G. Haghighi, Terrence J. Sejnowski, M. Margarita Behrens, and Joseph R. Ecker. Global epigenomic reconfiguration during mammalian brain development. *Science*, Jul 2013.

[3] Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J. Ziller, Viren Amin, John W. Whitaker, Matthew D. Schultz, Lucas D. Ward, Abhishek Sarkar, Gerald Quon, Richard S. Sandstrom, Matthew L. Eaton, Yi-Chieh Wu, Andreas R. Pfenning, Xinchen Wang, Melina Claussnitzer, Yaping Liu, Cristian Coarfa, R. Alan Harris, Noam Shoresh, Charles B. Epstein, Elizabeta Gjoneska, Danny Leung, Wei Xie, R. David Hawkins, Ryan Lister, Chibo Hong, Philippe Gascard, Andrew J. Mungall, Richard Moore, Eric Chuah, Angela Tam, Theresa K. Canfield, R. Scott Hansen, Rajinder Kaul, Peter J. Sabo, Mukul S. Bansal, Annaick Carles, Jesse R. Dixon, Kai-How Farh, Soheil Feizi, Rosa Karlic, Ah-Ram Kim, Ashwinikumar Kulkarni, Daofeng Li, Rebecca Lowdon, GiNell Elliott, Tim R. Mercer, Shane J. Neph, Vitor Onuchic, Paz Polak, Nisha Rajagopal, Pradipta Ray, Richard C. Sallari, Kyle T. Siebenthall, Nicholas A. Sinnott-Armstrong, Michael Stevens, Robert E. Thurman, Jie Wu, Bo Zhang, Xin Zhou, Arthur E. Beaudet, Laurie A. Boyer, Philip L. De Jager, Peggy J. Farnham, Susan J. Fisher, David Haussler, Steven J. M. Jones, Wei Li, Marco A. Marra, Michael T. McManus, Shamil Sunyaev, James A. Thomson, Thea D. Tlsty, Li-Huei Tsai, Wei Wang, Robert A. Waterland, Michael Q. Zhang, Lisa H. Chadwick, Bradley E. Bernstein, Joseph F. Costello, Joseph R. Ecker, Martin Hirst, Alexander Meissner, Aleksandar Milosavljevic, Bing Ren, John A. Stamatoyannopoulos, Ting Wang, Manolis Kellis, Andreas Pfenning, Melina ClaussnitzerYaping Liu, R. Alan Harris, R. David Hawkins, R. Scott Hansen, Nezar Abdennur, Mazhar Adli, Martin Akerman, Luis Barrera, Jessica Antosiewicz-Bourget, Tracy Ballinger, Michael J. Barnes, Daniel Bates, Robert J. A. Bell, David A. Bennett,

Katherine Bianco, Christoph Bock, Patrick Boyle, Jan Brinchmann, Pedro Caballero-Campo, Raymond Camahort, Marlene J. Carrasco-Alfonso, Timothy Charnecki, Huaming Chen, Zhao Chen, Jeffrey B. Cheng, Stephanie Cho, Andy Chu, Wen-Yu Chung, Chad Cowan, Qixia Athena Deng, Vikram Deshpande, Morgan Diegel, Bo Ding, Timothy Durham, Lorigail Echipare, Lee Edsall, David Flowers, Olga Genbacev-Krtolica, Casey Gifford, Shawn Gillespie, Erika Giste, Ian A. Glass, Andreas Gnirke, Matthew Gormley, Hongcang Gu, Junchen Gu, David A. Hafler, Matthew J. Hangauer, Manoj Hariharan, Meital Hatan, Eric Haugen, Yupeng He, Shelly Heimfeld, Sarah Herlofsen, Zhonggang Hou, Richard Humbert, Robbyn Issner, Andrew R. Jackson, Haiyang Jia, Peng Jiang, Audra K. Johnson, Theresa Kadlecek, Baljit Kamoh, Mirhan Kapidzic, Jim Kent, Audrey Kim, Markus Kleinewietfeld, Sarit Klugman, Jayanth Krishnan, Samantha Kuan, Tanya Kutyavin, Ah-Young Lee, Kristen Lee, Jian Li, Nan Li, Yan Li, Keith L. Ligon, Shin Lin, Yiing Lin, Jie Liu, Yuxuan Liu, C. John Luckey, Yussanne P. Ma, Cecile Maire, Alexander Marson, John S. Mattick, Michael Mayo, Michael McMaster, Hayden Metsky, Tarjei Mikkelsen, Diane Miller, Mohammad Miri, Eran Mukame, Raman P. Nagarajan, Fidencio Neri, Joseph Nery, Tung Nguyen, Henriette O'Geen, Sameer Paithankar, Thalia Papayannopoulou, Mattia Pelizzola, Patrick Plettner, Nicholas E. Propson, Sriram Raghuraman, Brian J. Raney, Anthony Raubitschek, Alex P. Reynolds, Hunter Richards, Kevin Riehle, Paolo Rinaudo, Joshua F. Robinson, Nicole B. Rockweiler, Evan Rosen, Eric Rynes, Jacqueline Schein, Renee Sears, Terrence Sejnowski, Anthony Shafer, Li Shen, Robert Shoemaker, Mahvash Sigaroudinia, Igor Slukvin, Sandra Stehling-Sun, Ron Stewart, Sai Lakshmi Subramanian, Kran Suknuntha, Scott Swanson, Shulan Tian, Hannah Tilden, Linus Tsai, Mark Urich, Ian Vaughn, Jeff Vierstra, Shinny Vong, Ulrich Wagner, Hao Wang, Tao Wang, Yunfei Wang, Arthur Weiss, Holly Whitton, Andre Wildberg, Heather Witt, Kyoung-Jae Won, Mingchao Xie, Xiaoyun Xing, Iris Xu, Zhenyu Xuan, Zhen Ye, Chia-an Yen, Pengzhi Yu, Xian Zhang, Xiaolan Zhang, Jianxin Zhao, Yan Zhou, Jiang Zhu, Yun Zhu, and Steven Ziegler. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015. ISSN 0028-0836. doi: 10.1038/nature14248. URL http://www.ncbi.nlm.nih.gov/pubmed/25693563.

[4] Hendrik G Stunnenberg, Sergio Abrignani, David Adams, Melanie de Almeida, Lucia Altucci, Viren Amin, Ido Amit, Stylianos E Antonarakis, Samuel Aparicio, Takahiro Arima, Laura Arrigoni, Rob Arts, Vahid Asnafi, Manel Esteller, Jae-Bum Bae, Kevin Bassler, Stephan Beck, Benjamin Berkman, Bradley E Bernstein, Mikhail Bilenky, Adrian Bird, Christoph Bock, Bernhard Boehm, Guillaume Bourque, Charles E Breeze, Benedikt Brors, David Bujold, Oliver Burren, Marion J Bussemakers, Adam Butterworth, Elias Campo, Enrique Carrillo-de Santa-Pau, Lisa Chadwick, Kui Ming Chan,

Wei Chen, Tom H Cheung, Luca Chiapperino, Nak Hyen Choi, Ho-Ryun Chung, Laura Clarke, Joseph M Connors, Philippe Cronet, John Danesh, Manolis Dermitzakis, Gerard Drewes, Pawel Durek, Stephanie Dyke, Tomasz Dylag, Connie J Eaves, Peter Ebert, Roland Eils, Jürgen Eils, Catherine A Ennis, Tariq Enver, Elise A Feingold, Bärbel Felder, Anne Ferguson-Smith, Jude Fitzgibbon, Paul Flicek, Roger S.-Y. Foo, Peter Fraser, Mattia Frontini, Eileen Furlong, Sitanshu Gakkhar, Nina Gasparoni, Gilles Gasparoni, Daniel H Geschwind, Petar Glažar, Thomas Graf, Frank Grosveld, Xin-Yuan Guan, Roderic Guigo, Ivo G Gut, Alf Hamann, Bok-Ghee Han, R Alan Harris, Simon Heath, Kristian Helin, Jan G Hengstler, Alireza Heravi-Moussavi, Karl Herrup, Steven Hill, Jason A Hilton, Benjamin C Hitz, Bernhard Horsthemke, Ming Hu, Joo-Yeon Hwang, Nancy Y Ip, Takashi Ito, Biola-Maria Javierre, Sasa Jenko, Thomas Jenuwein, Yann Joly, Steven J M Jones, Yae Kanai, Hee Gyung Kang, Aly Karsan, Alexandra K Kiemer, Song Cheol Kim, Bong-Jo Kim, Hyeon-Hoe Kim, Hiroshi Kimura, Sarah Kinkley, Filippos Klironomos, In-Uk Koh, Myrto Kostadima, Christopher Kressler, Roman Kreuzhuber, Anshul Kundaje, Ralf Küppers, Carolyn Larabell, Paul Lasko, Mark Lathrop, Daniel H S Lee, Suman Lee, Hans Lehrach, Elsa Leitão, Thomas Lengauer, Åke Lernmark, R David Leslie, Gilberto K K Leung, Danny Leung, Markus Loeffler, Yussanne Ma, Antonello Mai, Thomas Manke, Eric R Marcotte, Marco A Marra, Joost H A Martens, Jose Ignacio Martin-Subero, Karen Maschke, Christoph Merten, Aleksandar Milosavljevic, Saverio Minucci, Totai Mitsuyama, Richard A Moore, Fabian Müller, Andrew J Mungall, Mihai G Netea, Karl Nordström, Irene Norstedt, Hiroaki Okae, Vitor Onuchic, Francis Ouellette, Willem Ouwehand, Massimiliano Pagani, Vera Pancaldi, Thomas Pap, Tomi Pastinen, Ronak Patel, Dirk S Paul, Michael J Pazin, Pier Giuseppe Pelicci, Anthony G Phillips, Julia Polansky, Bo Porse, J Andrew Pospisilik, Shyam Prabhakar, Dena C Procaccini, Andreas Radbruch, Nikolaus Rajewsky, Vardham Rakyan, Wolf Reik, Bing Ren, David Richardson, Andreas Richter, Daniel Rico, David J Roberts, Philip Rosenstiel, Mark Rothstein, Abdulrahman Salhab, Hiroyuki Sasaki, John S Satterlee, Sascha Sauer, Claudia Schacht, Florian Schmidt, Gerd Schmitz, Stefan Schreiber, Christopher Schröder, Dirk Schübeler, Joachim L Schultze, Ronald P Schulyer, Marcel Schulz, Martin Seifert, Katsuhiko Shirahige, Reiner Siebert, Thomas Sierocinski, Laura Siminoff, Anupam Sinha, Nicole Soranzo, Salvatore Spicuglia, Mikhail Spivakov, Christian Steidl, J Seth Strattan, Michael Stratton, Peter Südbeck, Hao Sun, Narumi Suzuki, Yutaka Suzuki, Amos Tanay, David Torrents, Frederick L Tyson, Thomas Ulas, Sebastian Ullrich, Toshikazu Ushijima, Alfonso Valencia, Edo Vellenga, Martin Vingron, Chris Wallace, Stefan Wallner, Jörn Walter, Huating Wang, Stephanie Weber, Nina Weiler, Andreas Weller, Andrew Weng, Steven Wilder, Sam M Wiseman, Angela R Wu, Zhenguo Wu, Jieyi Xiong, Yasuhiro Yamashita, Xinyi Yang, Desmond Y Yap, Kevin Y Yip, Stephen Yip, Jae-Il Yoo, Daniel Zerbino,

Gideon Zipprich, and Martin Hirst. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell*, 167(5):1145–1149, jun 2017. ISSN 0092-8674. doi: 10.1016/j.cell.2016.11.007. URL http://dx.doi.org/10.1016/j.cell.2016.11.007.

[5] Gary C. Hon, Nisha Rajagopal, Yin Shen, David F. McCleary, Feng Yue, My D. Dang, and Bing Ren. Epigenetic memory at embryonic enhancers identified in dna methylation maps from adult mouse tissues. *Nat Genet*, 45: 1198–1206, Oct 2013.

[6] Y. He and J. R. Ecker. Non-CG Methylation in the Human Genome. *Annu Rev Genomics Hum Genet*, 16:55–77, 2015.

[7] J. U. Guo, Y. Su, J. H. Shin, J. Shin, H. Li, B. Xie, C. Zhong, S. Hu, T. Le, G. Fan, H. Zhu, Q. Chang, Y. Gao, G. L. Ming, and H. Song. Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. *Nat. Neurosci.*, 17(2):215–222, Feb 2014.

[8] L. Chen, K. Chen, L. A. Lavery, S. A. Baker, C. A. Shaw, W. Li, and H. Y. Zoghbi. MeCP2 binds to non-CG methylated DNA as neurons mature, influencing transcription and the timing of onset for Rett syndrome. *Proc. Natl. Acad. Sci. U.S.A.*, 112(17):5509–5514, Apr 2015.

[9] Harrison W. Gabel, Benyam Kinde, Hume Stroud, Caitlin S. Gilbert, David a. Harmin, Nathaniel R. Kastan, Martin Hemberg, Daniel H. Ebert, and Michael E. Greenberg. Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. *Nature*, 2015. ISSN 0028-0836. doi: 10.1038/nature14319. URL http://www.nature.com/nature/journal/vaop/ncurrent/full/nature14319.html?WT.ec{_}id=NATURE-20150312{#}affil-auth.

[10] S. Lagger, J. C. Connelly, G. Schweikert, S. Webb, J. Selfridge, B. H. Ramsahoye, M. Yu, C. He, G. Sanguinetti, L. C. Sowers, M. D. Walkinshaw, and A. Bird. MeCP2 recognizes cytosine methylated tri-nucleotide and di-nucleotide sequences to tune transcription in the mammalian brain. *PLoS Genet.*, 13(5):e1006793, May 2017.