Title

GENOME-SCALE STUDIES OF DYNAMIC DNA METHYLATION IN MAMMALIAN BRAIN CELLS

Author

Keown, Christopher Lee

Publication Date

2018

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

GENOME-SCALE STUDIES OF DYNAMIC DNA METHYLATION IN MAMMALIAN BRAIN CELLS

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

in

Cognitive Science

by

Christopher L. Keown

Committee in charge:

Professor Eran A. Mukamel, Chair
Professor Andrea A. Chiba
Professor Paula A. Desplats
Professor Joseph R. Ecker
Professor Fred H. Gage
Professor Lara M. Rangel
Professor Terrence J. Sejnowski

2018

The dissertation of Christopher L. Keown is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

_____

_____

Chair

University of California San Diego

2018

DEDICATION

I would like to dedicate this to my mother and father, Ruth Ann and Joe Keown. They taught me right from wrong and made endless sacrifices so I could have the opportunities they didn't. They gave me the space to follow my dreams and have been by my side to encourage me along the way. I would also like to dedicate this to Ms. Purviance, my 10th grade English teacher, who believed in challenging her students, even if that meant copious amounts of grading. You took the time to care and challenged me to go beyond my limits.

EPIGRAPH

It doesn't matter how beautiful your theory is,

it doesn't matter how smart you are.

If it doesn't agree with experiment, it's wrong.

—Richard Feyman

You can't even begin to understand biology, you can't understand life,

unless you understand what it's all there for, how it arose

–and that means evolution.

—Richard Dawkins

I want to put a ding in the universe.

—Steve Jobs

TABLE OF CONTENTS

Signature Page iii

Dedication iv

Epigraph v

Table of Contents vi

List of Figures x

List of Tables xii

Acknowledgements xiii

Vita xiv

Abstract of the Dissertation xvi

Chapter 1 Introduction 1
1.1 What are the effects of early life experience on DNA methylation and gene expression in the dentate gyrus of the hippocampus? 5
1.2 What is the role of neuron-specific non-CG DNA methylation in epigenetic silencing of a large chromatin domain during female X chromosome inactivation? 8
1.3 How does DNA methylation differ among the many types of excitatory and inhibitory neurons in the frontal cortex? 11

Chapter 2 Environmental enrichment increases transcriptional and epigenetic differentiation between mouse dorsal and ventral dentate gyrus 16
2.1 Introduction 16
2.2 Results 17
2.2.1 Environmental enrichment promotes hippocampal neurogenesis 17
2.2.2 Specialization of gene expression in dorsal and ventral DG . 18
2.2.3 DNA methylation differences between dorsal and ventral DG 20
2.2.4 DNA methylation correlates with repression at some genes . . 21
2.2.5 Impact of enrichment on DNA methylation in DG 22
2.2.6 *NeuroD1* binding sites enriched at dorsal DMRs 23
2.3 Discussion 24
2.4 Methods 28
2.4.1 Animals and environmental enrichment 28
2.4.2 Tissue collection for MRI 29
2.4.3 Magnetic resonance imaging and analysis 30

2.4.4 Immunohistochemistry . . . 31
2.4.5 Tissue collection for sequencing assays . . . 31
2.4.6 RNA and DNA extraction . . . 32
2.4.7 RNA-Seq collection . . . 32
2.4.8 Validation of RNA-Seq results by digital Nanostring . . . 32
2.4.9 TAB-Seq and MethylC-Seq . . . 32
2.4.10 Data analysis . . . 33
2.4.11 Differential expression . . . 34
2.4.12 Differential methylation analysis . . . 34
2.4.13 Gene ontology analysis . . . 36
2.4.14 Data availability . . . 36
2.5 Acknowledgements . . . 36

Chapter 3 Allele-specific non-CG DNA methylation marks domains of active chromatin in female mouse brain . . . 42
3.1 Introduction . . . 42
3.2 Results . . . 44
3.2.1 Allele-Specific Global Levels of CG and Non-CG DNA Methylation on Female X Chromosomes . . . 44
3.2.2 Differential Methylation Patterns at Genes Known to Escape XCI . . . 45
3.2.3 Differential mCH and mCG Between Xa and Xi Predict Escape Genes . . . 47
3.2.4 Analyses of Intergenic Regions . . . 48
3.2.5 Allele-Specific Methylation and Imprinting . . . 49
3.3 Discussion . . . 51
3.4 Methods . . . 54
3.4.1 Mouse Model/Animals . . . 54
3.4.2 MethylC-Seq . . . 55
3.4.3 mRNA-Seq Library Preparation . . . 55
3.4.4 Reference Genomes . . . 56
3.4.5 Mapping of MethylC-Seq Data . . . 56
3.4.6 Mapping of mRNA-Seq Data . . . 57
3.4.7 Data Analysis . . . 57
3.4.8 MethylC-Seq Analysis . . . 57
3.4.9 mRNA-Seq Analysis . . . 58
3.4.10 Definition of CH-Hypermethylated and CG-Hypomethylated Genes . . . 58
3.4.11 Additional Datasets . . . 59
3.4.12 Data Access and Browser . . . 59
3.5 Acknowledgements . . . 59

Chapter 4 Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex . . . 65
4.1 Article . . . 65
4.2 Acknowledgements . . . 72

Chapter 5 Conclusions . . . 77

Appendix A Supporting Information Chapter 2 . . . 80

Appendix B Supporting Information Chapter 3 . . . 86
B.1 SI Discussion: Validation of Allele-Specific Methylation Accuracy . 86

Appendix C Supporting Information Chapter 3 . . . 92
C.1 Materials and Methods . . . 92
C.1.1 Animal samples . . . 92
C.1.2 Human samples . . . 93
C.1.3 Mouse tissue dissections . . . 93
C.1.4 Nuclear isolation . . . 93
C.1.5 Flow cytometry based nuclei sorting . . . 93
C.1.6 Preparation of single nucleus methylome library . . . 94
C.1.7 Sequencing of single nucleus methylome library . . . 98
C.1.8 Single cell methylome mapping and data analysis . . . 98
C.1.9 MethylC-seq of mouse SST+ inhibitory neurons . . . 98
C.1.10 Genomic sequencing of the human sample . . . 99
C.1.11 Calling sequence variants for the human sample . . . 99
C.1.12 Data cleaning . . . 99
C.1.13 Clustering analysis . . . 100
C.1.14 Validation of clustering . . . 102
C.1.15 Processing of single cell and nucleus RNA-seq datasets . . . 104
C.1.16 Cross-species comparison of single neuron clusters . . . 105
C.1.17 Comparison of single cell clusters defined by single cell/nucleus RNA-seq and single cell methylome . . . 105
C.1.18 In situ hybridization (ISH) and image analysis . . . 105
C.1.19 Identification of CG-DMR and superenhancer-like large CG-DMRs . . . 106
C.1.20 Comparison of single cell methylome methods . . . 107
C.1.21 Transcription factor (TF) binding motif enrichment analysis 108
C.1.22 Prediction of putative enhancers . . . 108
C.1.23 Prediction of excitatory neuron super-enhancers . . . 110
C.1.24 Comparative analysis of regulatory elements . . . 110
C.2 Supplementary texts . . . 112
C.2.1 snmC-seq shows reliable sample multiplexing and high reads mapping rate . . . 112

C.2.2 hPv-2 is a potentially human specific PV+ inhibitory neuron population 113
C.2.3 Inhibitory neurons show layer-specific DNA methylation signatures 114
C.2.4 Large CG-DMR is a reliable marker for superenhancer . . . 115

Bibliography 133

LIST OF FIGURES

Figure 2.1: Transcriptional effects of enriched environment (EE) are greater in dorsal than ventral dentate gyrus (DG). 37
Figure 2.2: Reduced DNA methylation in dorsal dentate gyrus associated with expression differences. 38
Figure 2.3: Greater dorsal-ventral differentiation of DNA methylation in enriched environment. 39
Figure 2.4: Transcription factor binding sites are enriched at DMRs. 40
Figure 2.5: A model for epigenetic regulation of dorsal and ventral DG. 41

Figure 3.1: Ultrasparse mCH on Xi correlates with escape domains. 61
Figure 3.2: Escape genes are marked by unique mCG and mCH patterns. 62
Figure 3.3: X-chromosome-wide landscape of allele-specific DNA methylation. 63
Figure 3.4: Imprinted genes marked by allele-specific mCG and mCH. 64

Figure 4.1: High-throughput single-nucleus methylome sequencing (snmC-seq) of mouse and human frontal cortex (FC) neurons. 73
Figure 4.2: Non-CG methylation (mCH) signatures identify distinct neuron populations in mouse and human FC. 74
Figure 4.3: Conserved and divergent neuron type-specific gene regulatory elements. . . 75
Figure 4.4: Gene body mCH and CG-DMRs conserved between mouse and human. . . 76

Figure A.1: Increased adult neurogenesis in dorsal and ventral DG following environmental enrichment. 81
Figure A.2: Transcriptome analysis of dorsal and ventral DG in enriched environment. . 82
Figure A.3: Transcriptome principal components separate dorsal and ventral samples using mCG, mCH or 5hmCG. 83
Figure A.4: Browser views of examples of genes hypomethylated in ventral DG. 84
Figure A.5: Cross-validated analysis of the correlation of region- and treatment-based effects on DNA methylation. 85

Figure B.1: Topologically associated domains correlate with mCH. 88
Figure B.2: Asymmetrical expression and methylation distinguish Xa from Xi. 89
Figure B.3: Methylation on Xa and the male X chromosome are similar. 89
Figure B.4: *Bcor* shows diffuse CG hypomethylation on Xi, although it does not escape. 90
Figure B.5: Bidirectional allele-specific methylation at the Nesp/Gnas/Nespas locus. . . . 91

Figure C.1: mCH can be accurately estimated within 100kb bins using sparse snmC-seq data. 116
Figure C.2: snmC-seq is compatible with multiplexing and demonstrates efficient read mappability. 117
Figure C.3: Single nuclei are consistently clustered by cell type using multiple methylome features across a wide range of genomic length scales. 118

Figure C.4: Cluster robustness. 119
Figure C.5: Absence of strong association between neuron clusters and experimental factors. 120
Figure C.6: Mouse marker genes. 121
Figure C.7: Human marker genes. 122
Figure C.8: Single neuron clusters are correlated with layer dissection and bulk methylome generated from purified neuron populations. 123
Figure C.9: Correlation between single neuron clusters defined by snmC-seq and single cell/nucleus RNA-seq. 124
Figure C.10: Prediction of neuron type marker genes with single cell methylomes. . . . 125
Figure C.11: Double ISH experiments validate novel markers predicted by mCH. 126
Figure C.12: Expanded neuronal diversity in human FC. 127
Figure C.13: Global mC levels are conserved between mouse and human neuron types. . 128
Figure C.14: Inhibitory neurons possess global and gene level cortical-layer-specific mCH signatures. 129
Figure C.15: Neuron-type-specific CG-DMRs reveal regulatory diversity in human and mouse brains. 130
Figure C.16: Identification of neuron type specific large CG-DMRs with super-enhancer like properties. 131
Figure C.17: Regulatory conservation of neuron types 132

LIST OF TABLES

Table 3.1: Escape genes and CH-hypermethylated genes 60

ACKNOWLEDGEMENTS

Chapter 2, in full, is a reprint of the material as it appears in *Nature Communications* 9 (1):298, 2018. Tie-Yuan Zhang*, Christopher L Keown*, Xianglan Wen, Junhao Li, Dulcie A. Vousden, Christoph Anacker, Urvashi Bhattacharyya, Richard Ryan, Josie Diorio, Nicholas O'Toole, Jason P. Lerch, Eran A. Mukamel, Michael J. Meaney. The dissertation author was the co-investigator and co-first 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 of the United States of America*, https://doi.org/10.1073/pnas.1611905114, 2017. Christopher L. Keown, Joel B. Berletch, Rosa Castanon, Joseph R. Nery, Christine M. Disteche, Joseph R. Ecker, and Eran A. Mukamel. The dissertation author was the primary investigator and first author of this paper.

Chapter 4, in full, is a reprint of the material as it appears in *Science* 357 (6351): 600-604, 2017. Chongyuan Luo*, Christopher L. Keown*, Laurie Kurihara, Jingtian Zhou, Yupeng He, Junhao Li, Rosa Castanon, Jacinta Lucero, Joseph R. Nery, Justin P. Sandoval, Brian Bui, Terrence J. Sejnowski, Timothy T. Harkins, Eran A. Mukamel, M. Margarita Behrens, Joseph R. Ecker. The dissertation author was the co-investigator and co-first author of this paper.

VITA

2003	B.S. in Computer Science, Indiana University, Bloomington
2013	M.S. in Computational Science, San Diego State University
2013-2018	Graduate Research Assistant, University of California, San Diego
2018	Ph.D. in Cognitive Science, University of California, San Diego

PUBLICATIONS

Zhang, Tie-Yuan*, **Christopher L. Keown***, Xianglan Wen, Junhao Li, Dulcie A. Vousden, Christoph Anacker, Urvashi Bhattacharyya, et al. 2018. "Environmental Enrichment Increases Transcriptional and Epigenetic Differentiation between Mouse Dorsal and Ventral Dentate Gyrus." *Nature Communications* 9 (1):298.

Luo, Chongyuan*, **Christopher L. Keown***, Laurie Kurihara, Jingtian Zhou, Yupeng He, Junhao Li, Rosa Castanon, et al. 2017. "Single-Cell Methylomes Identify Neuronal Subtypes and Regulatory Elements in Mammalian Cortex." *Science* 357 (6351): 600-604.

Keown, Christopher L., Joel B. Berletch, Rosa Castanon, Joseph R. Nery, Christine M. Disteche, Joseph R. Ecker, and Eran A. Mukamel. 2017. "Allele-Specific Non-CG DNA Methylation Marks Domains of Active Chromatin in Female Mouse Brain." *Proceedings of the National Academy of Sciences of the United States of America*, March. https://doi.org/10.1073/pnas.1611905114.

Abbott, Angela E., Annika Linke, Aarti Nair, Afrooz Jahedi, Laura A. Alba, **Christopher L. Keown**, Inna Fishman, and Ralph-Axel Müller. 2017. "Repetitive Behaviors in Autism Are Linked to Imbalance of Corticostriatal connectivity: A Functional Connectivity MRI Study." *Social Cognitive and Affective Neuroscience*, November. https://doi.org/10.1093/scan/nsx129.

Keown, Christopher L., Michael C. Datko, Colleen P. Chen, Jose Omar Maximo, Afrooz Jahedi, and Ralph-Axel Müller. 2017. "Network Organization Is Globally Atypical in Autism: A Graph Theory Study of Intrinsic Functional Connectivity." *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 2 (1). Elsevier: 66-75.

Shen, Mark D., Deana D. Li, **Christopher L. Keown**, Aaron Lee, Ryan T. Johnson, Kathleen Angkustsiri, Sally J. Rogers, Ralph-Axel Müller, David G. Amaral, and Christine Wu Nordahl. 2016. "Functional Connectivity of the Amygdala Is Disrupted in Preschool-Aged Children With Autism Spectrum Disorder." *Journal of the American Academy of Child and Adolescent Psychiatry* 55 (9): 817-24.

Chen, Colleen P., **Christopher L. Keown**, Afrooz Jahedi, Aarti Nair, Mark E. Pflieger, Barbara A. Bailey, and Ralph-Axel Müller. 2015. "Diagnostic Classification of Intrinsic Functional Connectivity Highlights Somatosensory, Default Mode, and Visual Regions in Autism." *NeuroImage. Clinical* 8 (April): 238-45.

Khan, Amanda J., Aarti Nair, **Christopher L. Keown**, Michael C. Datko, Alan J. Lincoln, and Ralph-Axel Müller. 2015. "Cerebro-Cerebellar Resting-State Functional Connectivity in Children and Adolescents with Autism Spectrum Disorder." *Biological Psychiatry* 78 (9): 625-34.

Abbott, Angela E., Aarti Nair, **Christopher L. Keown**, Michael Datko, Afrooz Jahedi, Inna Fishman, and Ralph-Axel Müller. 2016. "Patterns of Atypical Functional Connectivity and Behavioral Links in Autism Differ Between Default, Salience, and Executive Networks." *Cerebral Cortex* 26 (10): 4034-45.

So, Alex Yick-Lun, Reeshelle Sookram, Aadel A. Chaudhuri, Aarathi Minisandram, David Cheng, Catherine Xie, Ee Lyn Lim, . . . , **Christopher L. Keown**, et al. 2014. "Dual Mechanisms by Which miR-125b Represses IRF4 to Induce Myeloid and B-Cell Leukemias." *Blood* 124 (9): 1502-12.

Fishman, Inna, **Christopher L. Keown**, Alan J. Lincoln, Jaime A. Pineda, and Ralph-Axel Müller. 2014. "Atypical Cross Talk between Mentalizing and Mirror Neuron Networks in Autism Spectrum Disorder." *JAMA Psychiatry* 71 (7): 751-60.

Nair, Aarti*, **Christopher L. Keown***, Michael Datko, Patricia Shih, Brandon Keehn, and Ralph-Axel Müller. 2014. "Impact of Methodological Variables on Functional Connectivity Findings in Autism Spectrum Disorders." *Human Brain Mapping* 35 (8): 4035-48.

Keown, Christopher Lee, Patricia Shih, Aarti Nair, Nick Peterson, Mark Edward Mulvey, and Ralph-Axel Müller. 2013. "Local Functional Overconnectivity in Posterior Brain Regions Is Associated with Symptom Severity in Autism Spectrum Disorders." *Cell Reports* 5 (3): 567-72.

Maximo, Jose O., **Christopher L. Keown**, Aarti Nair, and Ralph-Axel Müller. 2013. "Approaches to Local Connectivity in Autism Using Resting State Functional Connectivity MRI." *Frontiers in Human Neuroscience* 7 (October): 605.

Chen, Colleen Pam, **Christopher L. Keown**, and Ralph-Axel Müller. 2013. "Towards Understanding Autism Risk Factors: a Classification of Brain Images with Support Vector Machines" *International Journal of Semantic Computing* 07 (02): 205-13.

Di Martino, A., C-G Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson, . . . , **Christopher L. Keown** et al. 2014. "The Autism Brain Imaging Data Exchange: Towards a Large-Scale Evaluation of the Intrinsic Brain Architecture in Autism." *Molecular Psychiatry* 19 (6): 659-67.

Shih, Patricia, Brandon Keehn, Jessica K. Oram, Kelly M. Leyden, **Christopher L. Keown**, and Ralph-Axel Müller. 2011. "Functional Differentiation of Posterior Superior Temporal Sulcus in Autism: A Functional Connectivity Magnetic Resonance Imaging Study." *Biological Psychiatry* 70 (3): 270-77.

ABSTRACT OF THE DISSERTATION

GENOME-SCALE STUDIES OF DYNAMIC DNA METHYLATION IN MAMMALIAN BRAIN CELLS

by

Christopher L. Keown

Doctor of Philosophy in Cognitive Science

University of California San Diego, 2018

Professor Eran A. Mukamel, Chair

Developmental processes, genes and environmental factors interact to produce changes in cognition and behavior over the lifespan of an individual. However, the underlying molecular genetic mechanisms that mediate these changes remain to be fully elucidated. DNA methylation is an epigenetic mechanism that defines cell identity and helps regulate gene expression. DNA methylation is dynamic over development and has been shown to mediate experience-dependent changes, including those resulting from learning and memory and early life adversity. Although methylation mainly occurs at genomic cytosines in the CG dinucleotide context, methylation at non-CG sites was recently found in brain tissue. Non-CG methylation is specifically enriched in

neurons and accumulates during the early childhood stages of brain development. The biological impact of non-CG methylation in regulating gene expression and regulating cellular function, if any, remains unclear. A major challenge for addressing this question is the complexity and scale of the DNA methylation landscape, which includes nearly one billion cytosines throughout the genome that are potential sites of modification in every cell. Targeted studies of specific candidate genes and genomic loci do not elucidate the overall configuration of the cellular epigenome. Techniques for comprehensively mapping the genome-wide distribution of DNA methylation are powerful, but they require sophisticated new computational methods of analysis that can reliably distinguish and statistically validate epigenomic differences related to developmental and environmental factors.

In this thesis we develop new approaches to comprehensively analyze DNA methylation throughout the genome and with single base resolution in order to better characterize the role of CG methylation and elucidate the potential role of CH methylation in mammalian brain cells. First, we consider the impact of enriched early life (peri-pubertal) experience on DNA methylation and gene expression in the dentate gyrus of the hippocampus. In addition to its role in experience-dependent gene regulation, DNA methylation also plays a key role in innate developmental processes, including female X chromosome inactivation. We provide the first allele-specific DNA methylomes from the active and inactive X chromosomes in female brain, and use comprehensive genomic analyses to gain insight into the functional relationship between allele-specific DNA methylation and transcription. These two studies provide new evidence of the dynamic changes in DNA methylation in whole brain tissue caused by environmental and innate developmental factors. However, they do not address the heterogeneity of brain cell types, a hallmark of mammalian brain organization. To address the role of DNA methylation in brain cell diversity, we develop computational methods to analyze data from a new assay that measures single cell methylomes. Using these data, we show that brain cell methylomes can be clustered and used to assess neuronal heterogeneity in the frontal cortex of mouse and human. Upon

clustering cells, we are able to gain insight into the role of methylation in the establishment and maintenance of cellular identity in neuronal types.

Overall, this thesis adds to the increasing evidence that DNA methylation is a cell type-specific, dynamic epigenomic modification of brain cells that is impacted by, and may in turn help to regulate, neuronal development and adaptation. In addition, this thesis provides new computational methods for analyzing large-scale, whole-genome DNA methylation data sets and demonstrates their use in uncovering new insights into the mammalian brain epigenome.

Chapter 1

Introduction

The inception of the field of cognitive science resulted from efforts to synthesize information and approaches across disparate fields studying the same object, the brain. At one end of the spectrum, neuroanatomists had largely focused on describing the physical structure, and more recently, the cellular and molecular architecture of the brain. At the other end, fields including behavioral psychology had focused on studying behavior without necessarily grounding their theories in considerations of the physical brain. At the same time, advances in molecular biology, genetics, and genomics have revealed the critical roles and, in some cases, the mechanisms by which genetic and environmental factors influence cognition. Today, cognitive scientists taking advantage of recent progress in high-throughput genome sequencing have the opportunity to produce a new, multi-scale understanding of the key neurobiological processes that shape the brain and support cognitive function. A healthy brain is the product of an extended process of development, the process that shapes the physical brain and underpins a well-defined emergence of specific cognitive functions. For example, children under the age 3.5 do not have the ability to deceive others, whereas adolescents and adults possess this cognitive ability [1]. At the same time, cognition relies on plastic changes to the brain resulting from experiences that can modify the brain and lead to memory and learning. The role of genetic and molecular mechanisms in

mediating developmental and environmental changes in cognition remain unclear.

In this dissertation, I focus on the potential role of epigenetic mechanisms in mediating changes in the brain and adapting cognitive behavior. Formally, epigenetics is the study of mitotically heritable changes in gene function not resulting from changes in DNA sequence. Epigenetics generally includes covalent modifications of DNA and post-translational modification of amino acids on the amino-terminal tail of histone proteins, known as histone modifications [2]. Here, I focus on DNA methylation, which is a covalent modification of a genomic cytosine by the addition of a methyl group to form 5-methylcytosine. DNA methylation is generally hypothesized to suppress gene expression because of its overall negative correlation with transcription across the genome; however, recent work suggests this view is overly simplistic and may need to be revised [3]. Because DNA methylation is a stable modification to the genome that modulates transcription, it could serve as a mechanism for mediating biological changes in neurons that underpin developmental and environmentally-induced changes in cognition. Previous work has indeed shown a global remodeling of DNA methylation in neurons during early childhood and adolescence, coincident with synaptogenesis [4]. Furthermore, neuron types in the adult brain have distinct methylation profiles that correlate with cell type-specific transcriptional profiles [5]. In addition, previous work suggests that environmental factors during development, such as nursing behavior in a mother rat, can alter DNA methylation in offspring that ultimately leads to behavioral changes [6, 7, 8]. However, additional work is required to better understand the role of DNA methylation in development and plasticity.

In mammals, DNA methylation primarily occurs at cytosine nucleotides followed by guanines, referred to as CpG or CG sites. However, recent work has uncovered an abundance of methylation at cytosines followed by bases other than guanine (non-CG sites, denoted CH) in neurons [9, 4] and in embryonic stem cells [10, 11]. Non-CG methylation (mCH) is negatively correlated with transcription in neurons, yet positively correlated with expression in stem cells, suggesting its function is highly contingent upon its interaction with cell type-specific factors.

Whereas developmental changes in CG methylation occur primarily at gene promoters and distal regulatory regions such as enhancers, non-CG methylation has a cell type-specific distribution throughout the whole gene body of many genes. mCH accrues in neurons over development coincident with synaptic proliferation and shows distinct profiles between excitatory, VIP+ and PV+ inhibitory neurons [5]. However, the potential function of mCH in modulating transcription factor binding and regulating gene expression remains to be characterized. Changes in CG methylation may mediate the impact of environmental factors on brain function [12], but mCH has not been examined in this context.

A broad range of assays have been developed for measuring DNA methylation with different trade-offs between resolution, breadth of the genome measured, ability to distinguish between CG and non-CG methylation, and cost. A primary goal of this dissertation was to measure non-CG methylation, which required base-resolution assays that could distinguish methylcytosines occuring in different local sequence contexts. Micro-array approaches, such as the Infinium Human Methylation 450K BeadChip and Infinium MethylationEPIC array, contain probes for a substantial fraction of the 25 million CG sites throughout the genome. However, these platforms contain only a few thousand probes for non-CG sites, which cannot adequately represent the $\sim$950 million CH positions in the genome. Immunoprecipitation-based assays such as methylated-DNA immunoprecipitation followed by sequencing (MeDIP-Seq) [13] cannot discriminate between CG and non-CG methylation, and their affinity for non-CG methylation in general remains unclear. Bisulfite sequencing (methylC-seq) allows for the distinct measurement of CG and non-CG methylation with a very low rate of error (typically $< 0.5\%$ false positives) [14, 15]. In methylC-Seq, DNA is treated with sodium bisulfite, converting unmethylated cytosine to uracil but leaving methylcytosine intact. Bisulfite-treated DNA fragments are then subjected to next generation sequencing. Computational procedures map the fragments to their genomic position through comparison to a reference genome, and methylated cytosines can be identified. Both the quality and the cost of methylC-Seq data are proportional to the amount of sequencing,

as each sequenced fragment represents an independent measurement of the methylation status of a particular set of positions in a particular cell. Targeted methylC-Seq and reduced representation bisulfite sequencing (RRBS) [16] allow specific regions of the genome to be selected and assayed with high coverage. Here, however, we were interested in broadly characterizing non-CG methylation across the whole genome to better understand the landscape of non-CG methylation. Therefore, we focused on data from whole genome methylC-Seq (WGBS), the gold standard for measurement of DNA methylation [15].

A limitation of WGBS is that genome-wide data sets have limited sequencing depth (number of mapped reads) at each genomic location. WGBS is also suited to population scale analysis, and data sets are typically limited to a handful of biological replicates. These considerations reduce the statistical power of WGBS data, while at the same time the large number of statistical tests required to examine the whole genome creates a challenging problem of multiple comparisons. Despite these challenges, WGBS has been successfully used to find hundreds of thousands of sites throughout the genome, called differentially methylated regions (DMRs), where DNA methylation is significantly different across tissues, cell types, and developmental ages. Optimal computational and statistical procedures for detecting and characterizing epigenomic differences using WGBS data are still a target of active research. This thesis contributes to the field by developing new analysis procedures for WGBS data and demonstrating how they may be used to address three key questions. In what follows, I introduce each of the three research questions that motivate the three main chapters of the thesis. In each case, I will highlight the specific computational analysis challenges that my work addressed to enable our research findings.

1.1 What are the effects of early life experience on DNA methylation and gene expression in the dentate gyrus of the hippocampus?

My research here begins by first examining the role of DNA methylation in mediating environmental effects on cognition and behavior. Early life adversity (ELA), in which an individual experiences a traumatic event early in life during development, is strongly associated with negative mental health outcomes in adults [7]. Because DNA methylation is dynamic during development yet largely stable in adults, it is an appealing mechanism for mediating long-lasting transcriptional changes in ELA. For example, a high level of maternal care during early life in rats, as defined by a large amount of the mother's time spent on licking, grooming and arched-back nursing, is associated with reduced fear and an attenuated stress response by the HPA axis later in life. Using maternal care as a model of ELA, Weaver et al. examined if DNA methylation could mediate the aforementioned behavioral changes [6]. These authors measured methylation in the hippocampus at the promoter of the glucocorticoid receptor (GR) gene, *Nr3c1*, whose expression dampens the hypothalamic-pituitary-adrenal (HPA) axis response to stress [17]. Their results showed increased methylation at the GR gene promoter in the stressed group that emerged over development and was retained into adulthood.

These studies catalyzed research into behavioral epigenetics and inspired many additional experiments. Changes in methylation at the GR gene promoter resulting from ELA have been identified in mice and humans, and additional research suggests GR promoter methylation is also modified in offspring of mothers stressed during pregnancy in both mouse and humans [7]. In a separate but complementary line of work, researchers have identified changes in methylation at synaptic plasticity genes that may mediate memory formation [18, 19]. Although this research has identified experience-dependent changes in DNA methylation that correlate with behavioral

and cognitive consequences, additional research is required to better understand the scope of these changes and their role in other behavioral contexts.

To address this challenge, I use WGBS in Chapter 2 of my dissertation to comprehensively analyze the effects of early life experience on DNA methylation and gene expression in the dentate gyrus of the hippocampus. The experimental work for this project was carried out by Tie-Yuan Zhang, Michael Meaney and their colleagues at McGill University, while I was primarily responsible for computational analysis and interpretation of the large-scale epigenomic and transcriptomic data. Our study examined mice raised in an enrichmented environment (EE), an experimental paradigm in which animals are exposed to a rich and dynamic environment during development, starting on postnatal day 22. In contrast to ELA, EE confers a positive outcome in adults and may confer a resilience to stress [20], making it of clinical interest. EE can induce dramatic changes in the brain and behavior in rodents, including enhanced synaptogenesis and cognitive improvements in spatial memory, working memory, and contextual fear memory [21]. Furthermore, EE leads to differential expression of neuronal activity and synaptic plasticity genes [22, **?**], as well as adult neurogenesis in the dentate gyrus (DG)[23]. The mechanisms that mediate these effects, however, are not fully understood. Just as DNA methylation may play a role in mediating the negative effects of ELA, it may also provide a mechanism for mediating the positive effects of EE in DG of the hippocampus.

In contrast with prior studies focusing on specific candidate genes, we aimed to comprehensively assess the effect of EE on the landscape of DNA methylation and gene expression throughout the genome. Because the effects of EE on individual animals are variable, we used five cohorts of mice in each condition (EE or SH) to generate independent biological replicates of DNA methylome (WGBS) and transcriptome (RNA-Seq) data. We separately examined the dorsal and ventral poles (dDG and vDG, respectively) because despite their molecular, functional and connectivity differences [24], their relative contributions to EE remain unknown. Comprehensive epigenomic and transcriptomic mapping is virtually unprecedented at this scale in studies of

behavior and gene regulation. To take full advantage of the power of these data, we could not rely on existing bioinformatic techniques for statistical analysis of differential expression and differential methylation that perform separate comparisons of each gene or genomic region. Instead, I designed and implemented a new approach to analyze differential methylation across many regions of the genome, and to relate these to the activity of specific DNA-binding transcription factors.

Consistent with previous literature, we found large transcriptional changes between the dorsal and ventral DG including upregulation of neurogenesis-associated genes in the dorsal region. DNA methylation also showed large regional differences, including at developmental patterning genes such as *Nr2f1* and *Nr2f2*. I observed a large asymmetry in regional DMRs, with most hypomethylation occuring in the dorsal DG. Comparison of EE and SH showed smaller, although pronounced, transcriptional differences. Although detecting differences in DNA methylation between EE and SH treated animals was one of our main aims, standard bioinformatic analyses identified a very small set of differentially methylated regions (DMRs) which were not clearly interpretable. I therefore developed a new computational method, which took advantage of the clearly identifiable DMRs from our comparison of dorsal and ventral DG to perform a focused analysis of EE-dependent differences in methylation at those regions. Comparison of methylation at regional DMRs between EE and SH showed increased hypomethylation in the dorsal region. Importantly, these DMRs were enriched enriched for the binding motif of NeuroD1, an important neurodevelopmental factor in adult neurogenesis, which may play a role in mediating the effects of EE.

1.2 What is the role of neuron-specific non-CG DNA methylation in epigenetic silencing of a large chromatin domain during female X chromosome inactivation?

My dissertation work was largely motivated by the discovery of non-CG methylation in neurons [4] and aimed at understanding the functional significance, if any, of this unique epigenetic marker. Therefore, it was important to study the effects of mCH in a well-characterized system of gene regulation that could provide clear insight while minimizing the number of unknown variables. The process of X chromosome inactivation (XCI) provides a well-studied model system, and therefore, we examined the connection between mCH profiles and transcription in neurons under XCI in Chapter 3.

Females have two copies of the X chromosome, whereas males only have one. XCI is the process in mammals that inactivates one of the two X chromosomes in each cell in females to avoid a potentially deleterious difference in gene dosage for the nearly $\sim$1,000 X-linked genes [25]. In many species, including humans and mice, the inactivated allele is selected during early development at random and independently across cells, thus leading to a mosaic pattern of inactivation. Importantly, some genes on the inactive X chromosome escape the process of inactivation and remain expressed from the inactivated allele. The proportion of genes escaping XCI is estimated to be around 3% of X-chromosome genes in mouse and 15% in humans [26, 27]. These escape genes are hypothesized to be critical in healthy brain development. For example, females with Turner syndrome (45,X), who have only a single X chromosome, lack the inactive allele (Xi) and thus escape genes are not expressed. Turner syndrome individuals have ADHD-like symptoms and nonverbal learning disabilities [28]. Thus the mechanisms that modulate expression on the inactive X chromosome (Xi) may help us understand how healthy brain development unfolds and also how sex differences in cognition arise.

Sharp et al. compared CG methylation on the X chromosome in human blood between Turner syndrome and typical females [29]. Promoter mCG was reduced at escape genes yet increased at genes that do not escape XCI, consistent with the role of promoter mCG as a transcriptional repressor. In their paper on the discovery of non-CG methylation in neurons, Lister et al. compared non-CG methylation in gene bodies on the X chromosome between females (aggregate levels on Xa and Xi) and males [4]. They found a 50% reduction of mCH on the X chromosome in females relative to males, except at escape genes, which show increased mCH in females. However, the stochastic nature of X-inactivation prevented the group from distinguishing between methylation on Xa and Xi explicitly. Thus, additional work is required to control for the stochasticity of XCI and to identify the source of the methylation and its associated functional significance.

To address this question, we used a model of deterministic X-inactivation [30] to measure levels of mCG and mCH on Xa and Xi separately. Through a collaboration with Christine Disteche and Joel Berletch (University of Washington), we obtained samples of frontal cortex from mouse F1 offspring resulting from a cross of a wild-type father and a mutant mother carrying a deletion in the XCI-initiating long non-coding RNA, Xist [30]. The paternal allele in the heterozygous offspring animals is always inactivated, thus circumventing the difficulties of studying random XCI. In addition, the mother and father are from different mouse species, C57/Bl6 maternal and Mus. spretus paternal, whose genomes differ at approximately 42 million single nucleotides. These genetic variants allowed us to assign most sequencing reads back to the allele of origin, thereby obtaining mCG and mCH for the maternal X (Xa) and the paternal X (Xi). For this experiment, we collected tissue from the frontal cortex of two replicate mice at eight weeks of age (in collaboration with Christine Disteche and Joel Berletch at University of Washington) and performed whole genome bisulfite sequencing to measure DNA methylation and RNA-seq (in collaboration with the Joseph R. Ecker and Rosa Castanon at The Salk Institute for Biological Sciences).

A major computational challenge in this project was the need for allele-specific analysis of DNA methylome (WGBS) and transcriptome (RNA-Seq) data. Standard bioinformatic approaches do not distinguish the two alleles, obscuring the allele-specific differences in gene expression and DNA methylation that we aimed to observe. Allele-specific analysis can take advantage of genetic variants between the two parental strains (heterozygous sites) to map sequencing fragments to their parent of origin. At the same time, we had to ensure that our analysis did not erroneously infer differences in DNA methylation due to what were, in fact, genetic differences between the mouse strains used in our study. Following the model of a previous allele-specific study in a different mouse model [9], I developed and validated a bioinformatic pipeline that accurately assessed DNA methylation and gene expression on each allele.

Our analysis identified an intriguing and complex relationship between DNA methylation and XCI in the brain. mCG recapitulated the patterns previously observed in blood [29]: mCG was high throughout Xi except at the promoters of escape genes, where it was depleted. In stark contrast to mCG, however, mCH was largely absent throughout Xi except at the gene bodies of escape genes, where it was highly enriched. We hypothesized that this difference between mCG and mCH may result from the timing of their appearance with respect to when XCI occurs. CG methylation is present before XCI occurs and can thus be utilized to help silence transcription on Xi. However, mCH in neurons does not appear until after the peak of expression of the DNA methyltransferase DNMT3A, around 1-2 weeks after birth in mice. By this time, XCI has already occurred, thus mCH does not accumulate because chromatin in Xi may be inaccessible to the methyltransferase. By contrast, escape genes reside in loops of open chromatin with a higher degree of accessibility, thus enabling DNMT3A to mediate mCH accumulation at these genes. Future work is required to examine this hypothesis directly.

1.3 How does DNA methylation differ among the many types of excitatory and inhibitory neurons in the frontal cortex?

In the final chapter of my dissertation, I focus on understanding how DNA methylation can help resolve the challenge of cellular heterogeneity in brain tissue. One challenge for tissue-based molecular biology research is resolving the underlying source of an effect of interest in heterogeneous tissue. Typically, brain tissue is dissected and analyzed as an aggregate of the underlying cells composing the tissue, and therefore the signal is dominated by majority cell types. For example, in our XCI project described above, the tissue was dissected from whole mouse frontal cortex, which comprises many types of neurons and glial cells. The signal in the RNA-seq and WGBS data is thus likely dominated by excitatory neurons and glial cells given their high abundance in the tissue, whereas inhibitory neurons will have a small contribution to the signal given their relatively low abundance. In the case when *a priori* knowledge suggests the effect of interest is specific to particular cell types, transgenic mouse lines coupled with flow cytometry-based cell sorting can be used to isolate the population interest (e.g., VIP+ neurons, SST+ neurons, etc.). Often, however, we do not have *a priori* information about the effect of interest and would like an unbiased approach to examine all cell types separately but in parallel. In other cases, genetic tools for targeting a particular cell population may not exist, making cell type specific analysis a challenge. In these situations, assaying many thousands of individual cells followed by computational analysis to cluster them based on similarity could provide information about all cell types. This insight has lead to the development of single cell technologies, such as single cell RNA-seq, which allows for measuring the transcriptome in individual cells [31, 32, 33, 34].

In Chapter 4, we develop and apply a single cell DNA methylation assay to characterize neuronal heterogeneity across thousands of neurons in the cortex of both mouse and human. This work was a a close collaboration with Chongyuan Luo and Joseph R. Ecker (Salk Institute) and Swift Biosciences, who developed and optimized the biochemical procedures for efficiently

performing bisulfite sequencing in single nuclei. The data were generated in collaboration with the neurobiology lab of M. Margarita Behrens (Salk Institute). My role in this project was to lead the computational analysis of large-scale single cell DNA methylome data sets. Although there have been several recent studies of single cell RNA-Seq data from which we could draw inspiration, there were no existing tools or procedures that were appropriate for analysis of single cell epigenomes.

Identification of neuron types is essential for understanding how the brain functions in typical individuals and how this goes awry in disorders such as autism and schizophrenia. Furthermore, understanding neuron type-specific disruption of molecular pathways in neurocognitive disorders can provide targets for their treatment. Neurons can be classified using attributes such as morphology, connectivity profiles, molecular markers, and physiological properties, such as axon potentials [35]. An exhaustive assessment of cell types requires a scalable approach that can randomly sample thousands or even millions of cells in a tissue to provide a sufficient quantity to detect minority cell types. Recent advances in the isolation of single cells or single nuclei allows the molecular marker approach to neuron type identification to achieve this scale, whereas anatomical and physiological approaches are currently more limited in their throughput. Several groups have examined neuronal heterogeneity using single-cell (scRNA-seq) or single-nucleus (snRNA-Seq) approaches to obtain transcriptomic profiles in individual cells. Zeisel et al. examined mouse somatosensory cortex and hippocampus and identified 47 cells types [36]. Tasic et al. examined visual cortex in mouse and identified 49 cell types including glia [37], and more recently examined two cortical regions, identifying 116 cell types [38]. In humans, Lake et al. examined six regions of the cortex and identified a total of 16 cell types [39]. Examining the transcriptome can give insight into short term changes of the molecular profile of cells. However, these approaches are highly sensitive to tissue handling and library preparation, and they may also be affected by factors such as circadian rhythms and activity-dependent changes in gene expression. Furthermore, the transcriptome represents $< 5\%$ of the genome, and other regions of

the genome may contain additional information essential for cell type identity.

Single cell DNA methylation can address the limitations of scRNA-seq approaches. First, whole genome methylomes assays, such as WGBS, measure methylation throughout the entire genome. Therefore, they can provide information about how cell identity is established and maintained, such as the location of enhancers and super-enhancers and their methylation status. Furthermore, DNA methylation is a stable covalent modification of DNA. Consequently, it provides a more stable representation of cell identity and is also less sensitive to alteration during sample processing and library preparation. Previous to our work, large-scale single cell analyses focused on RNA, and therefore, analytical methods were specific to the statistical properties of RNA data. snmC-seq, however, differs fundamentally in comparison to scRNA-seq data, requiring the development of novel analysis pipelines. scRNA-seq detects the presence of RNA transcripts and reports the total number of transcripts per gene. The data follows a Poisson or negative binomial distribution and cannot distinguish between a gene not being expressed and a gene not being measured due to so-called dropout. On the other hand, snmC-seq measures the presence of methylation at cytosines in the genome. The data is binomially distributed for some methylation level, p, and a number of trials, n (known as base calls). Unlike with scRNA-seq, snmC-seq can distinguish between absence of methylation (p=0, n >0) and the lack of measurement at a given cytosine ($n = 0$). In addition, a large portion of the genome is destroyed by bisulfite treatment, yielding an average of $\sim$5-10% coverage distributed randomly throughout the genome for each cell. Finally, snmC-seq data measures two types of methylation, mCG and mCH, which have differing statistical properties that must be carefully considered in our analyses. Given these unique characteristics of snmC-seq, we must develop custom processing and analysis pipelines.

Formal identification of cell types is an unsupervised clustering problem. Because we lack ground truth data for comparison and validation of our clustering results, we faced the challenge of devising unsupervised clustering techniques that identify biologically and statistically meaningful clusters without a clearly defined objective point of reference. On one hand, choosing parameters

that will completely divide all true clusters into their respective types can also lead to the appearance of false clusters that were driven by noise in the data, e.g. due to sparse coverage of the genome. On the other hand, under-clustering the data by grouping cells into a small set of overly broad cell types may also lead to inaccurate conclusions. Although it is difficult to identify the perfect clustering procedure, we relied on specific statistical and biological criteria to guide our choices and provide reproducible results. First, to ensure that our clusters were statistically valid (i.e. that they do not reflect technical noise), we performed clustering on a downsampling of cells and reads. We also performed clustering using differing types of clustering algorithms and compared their similarity. Second, to biologically validate our clusters we used an entropy measure to identify marker genes whose methylation was correlated with the clustering. These genes could then be compared to the literature and also to existing scRNA-seq datasets.

To address these challenges, we modified an existing scRNA-seq clustering approach called backSPIN to work on snmC-seq data [36]. BackSPIN is a divisive algorithm, where all cells start in the same cluster and are recursively divided into sub clusters, and thus provides a hierarchy for how the clusters were generated. We chose parameters that would initially over-cluster the cells. We then merged pairs of clusters that did not differ significantly at 7 or more marker genes (genes strongly hypomethylated [mCH in the bottom 2nd percentile] in one cluster and hypermethylated [mCH above the 80th percentile] for the other cluster). The next objective was to interpret our clusters against known characteristics of cortical neurons. We clustered our cells alongside WGBS data from purified VIP+, PV+, SST+, and Camk2a+ (pan-excitatory) populations in mouse in order to identify excitatory and specific inhibitory subpopulations. Cortical layer markers such as *Cux2*, *Rorb*, *Deptor*, and *Tle4* nicely corresponded with our clusters and were useful for discerning excitatory layers, whereas GABAergic markers, such as *Satb2*, *Gad1* and *Slc6a1*, further validated inhibitory and excitatory differences. In addition, for mouse the cortex was sliced into a top, middle and deep layer, and the cells were sequenced separately. These control layers strongly corresponded with the layers as predicted by

marker genes.

Our results identified 16 cell types in the mouse frontal cortex and 21 in human. We observed an expansion of the deep layers in both mouse and human. Whereas we only detected one layer 2/3 cluster and layer 4 cluster in mouse and human, we detected many more subtypes in the deep layers 5 and 6. This deep layer expansion was further increased in human compared to mouse. Although we would not expect a perfect match between scRNA-seq and snmC-seq, our neuron types were relatively consistent with previous scRNA-seq studies in both mouse and human [36, 37, 39]. Our results recapitulated numerous well-known marker genes and also identified new markers. We identified super-enhancers-like domains for each of the cell types, provided evidence for how the cell type identity is established and maintained. Finally, comparison of methylation in similar cell types between mouse and human show strong conservation of methylation and sequence in inhibitory neurons compared to excitatory neurons. Recent research has shown the excitatory neurons are different throughout the cortex, whereas inhibitory neurons are largely the same [38]. Therefore, inhibitory neurons may be generalists with more selective pressures and are thus constrained by all cortical regions, whereas excitatory neurons able to adapt and processing specific types of data in the cortex, and are thus under selective pressure to evolve.

Chapter 2

Environmental enrichment increases transcriptional and epigenetic differentiation between mouse dorsal and ventral dentate gyrus

2.1 Introduction

The hippocampus is implicated in learning and memory as well as the processing of emotional stimuli and regulation of stress responses. Dorsal and ventral hippocampal regions exhibit distinct connectivity and functional roles despite similar cell type composition [24]. The dorsal hippocampus, corresponding to the posterior hippocampus in primates, associates closely with cognitive functions and age-related cognitive impairments. In contrast, the ventral hippocampus, (anterior region in primates) is implicated in the regulation of emotional states and vulnerability for affective disorders. This functional specialization is reflected in patterns of gene expression. Gene expression in the dorsal hippocampus correlates with that in cortical

regions involved in information processing, while genes expressed in the ventral hippocampus correlate with expression in limbic regions involved in emotion and stress [24]. In addition, transcriptomic analysis reveals profound molecular differences, even within a uniform cell type population such as dorsal and ventral DG granule cells [40]. Epigenetic regulation may underlie these molecular differences and is also a potential mechanism for environmental influences on hippocampal development [41].

Early life experience has a profound, lifelong impact on emotional health due, in part, to environmental factors that influence gene expression in brain regions critical for cognitive-emotional stress responses. Epigenetic mechanisms such as DNA methylation, demethylation, and chromatin remodeling, have been linked to adult neurogenesis in the DG [42] and to neuronal plasticity underlying learning and memory [43, 44]. DNA methylation could likewise play a role in mediating long-term effects of early life experience [45]. Epigenetic modifications of DNA and histone proteins also define tissues and cell types during development [46, 5, 47], complicating the interpretation of epigenomic data from heterogeneous samples.

To elucidate the role of region-specific epigenetic regulation in the DG, we generated transcriptomes and base-resolution, whole-genome DNA methylation and hydroxymethylation profiles for the dorsal and ventral DG. Our data and analyses reveal substantial asymmetries between the DNA methylomes of the two hippocampal poles, and suggest that enriched environment (EE) enhances dorsal-specific epigenomic signatures.

2.2 Results

2.2.1 Environmental enrichment promotes hippocampal neurogenesis

Using high-resolution *in vivo* structural magnetic resonance imaging (MRI) [48, 49], we found that hippocampal volume is enlarged in mice raised in an enriched environment (EE) compared with standard housing (SH) in both the dorsal (8.5% greater volume, $p = 0.001$, t-

test) and ventral poles (6.1%, $p = 0.039$; significant interaction between region and condition, $p = 0.017$) (Fig. 2.1A). EE also associates with >60% more newborn neurons labeled by 5'-bromo-2'-dexoyuridine (BrdU), a marker of proliferating cells [50], in the DG (dorsal, $p = 0.0097$; ventral, $p = 0.028$; Appendix A.1A, B). These results are consistent with previous findings that enrichment increases hippocampal volume and neurogenesis in the dentate gyrus [48, 49].

2.2.2 Specialization of gene expression in dorsal and ventral DG

To address the molecular basis for the effect of EE on hippocampal function, we used RNA-Seq to profile gene expression in dorsal and ventral DG. Dentate granule cells have distinct gene expression patterns at the two poles [40], and single-nucleus transcriptome profiling has been used to link patterns of gene expression with the developmental trajectory of newborn neurons [51] and the activation of immediate early genes in a novel environment [34]. However, the impact of environmental enrichment on the specialized gene expression programs of the dorsal and ventral DG has not been examined. To increase the statistical power of our gene expression analysis and to limit variability due to single-nucleus isolation or microdissection, we performed RNA-Seq in carefully dissected whole-tissue samples of dorsal and ventral DG from 5 independent biological replicates in each condition (each replicate used pooled tissue from $n = 10 - 12$ animals; see STAR methods). Compared with microdissection-based RNA-Seq data [40], our gene expression profiles showed high correlation between samples (Spearman correlation for replicates, $r = 0.988$ compared with $r = 0.785$, Appendix A.2A-F). This level of quantitative precision in our data allowed us to comprehensively detect gene expression changes due to EE in the dorsal and ventral DG. Although our samples from whole tissue comprise multiple neuronal and glial cell types, the gene expression profiles we observed were most strongly correlated with expression from purified neurons compared to non-neuronal brain cell types, suggesting the tissue is primarily composed of neurons (Appendix A.2P) [52].

Transcriptome-wide analysis showed that dorsal-ventral differences in gene expression

account for nearly half of the total data variance (Fig. 2.1B, Appendix A.2G). Over 28% of genes expressed in the DG were differentially expressed by region (3,497 out of 12,247 genes; false discovery rate (FDR) < 0.05, TPM>1, fold-change $> 20\%$, Fig. 2.1C; Supplementary Data), including 244 genes (2%) with $>$2-fold difference in expression. Genes that were previously reported to show skewed expression in dorsal vs. ventral dentate granule cells [40] were similarly skewed in our data (Appendix A.2F), including dorsally enriched *Lct, Abcb10* and *Spata13* and ventrally enriched *Trhr*, *Grp*, and *Cpne7*. This consistency further supports the substantial contribution of granule neurons to our RNA-Seq data.

We found similar numbers of genes upregulated in the dorsal and the ventral regions. Although differential expression was widespread, the magnitude of expression differences was 4-fold smaller than the differences between distinct cortical cell types [5] (Appendix A.2H). We found notable differences between dorsal and ventral expression of key developmental factors such as ventrally-upregulated *Nr2f1/2* and dorsally-upregulated *Epha7*. Transcription factors that mark radial glia-like (RGL) stem cells (e.g. *Sox2*, *Hes5*) were enriched in the ventral DG, whereas maturing neuron markers (e.g., *NeuroD1*, *DCX*) were enriched in the dorsal DG (Fig. 2.1D,E) [42, 53], consistent with more active neurogenesis in the dorsal DG [54]. These data suggest specialized transcriptional regulation of neurogenesis in the dorsal and ventral DG.

Gene expression was more affected by EE in dorsal than ventral DG (Fig. 2.1B, greater separation of EE and SH samples on PC3 for dorsal than ventral, Appendix A.2I,J), and dorsal DG has twice as many differentially-expressed genes (152 dorsal, 72 ventral; FDR$<$0.05 and fold change $\geq 20\%$; Fig. 2.1F; Supplementary Data). The 37 genes upregulated in both regions were enriched for learning and memory function and included genes induced during neuronal activation (*Junb*, *Arc*, *Fos*, *Npas2/4*) that play critical roles in contextual memory formation [55]. *Gadd45b* was upregulated by EE in both regions and is implicated in activity-induced demethylation of gene promoters associated with neurogenesis [55]. Overall, our transcriptome analyses based on RNA sequencing, which we validated with amplification-free digital RNA quantification (Appendix

A.2K-N), are consistent with enhanced neurogenesis following EE, particularly in the dorsal DG.

2.2.3 DNA methylation differences between dorsal and ventral DG

DNA methylation is a stable epigenetic mark that could mechanistically support enduring transcriptional differences between dorsal and ventral DG and mediate the lifelong effects of early experience. Neuronal cell types show unique patterns of both CG and non-CG methylation (denoted mCG, mCH) [5], as well as hydroxymethylation (hmC) [56, 4]. However, methylation differences have not been examined within relatively homogenous cell types such as dentate granule cells arrayed along the longitudinal axis of the DG. Our RNA-Seq data showed that enzymes involved in DNA methylation (Dnmt1, Dnmt3a,b) and demethylation (Tet1,2,3, Gadd45a) are enriched in the dorsal compared to the ventral pole of the DG (Appendix A.2O). To examine mCG, mCH and hmC with single base resolution genome-wide, we performed bisulfite sequencing (MethylC-Seq) and Tet-assisted bisulfite sequencing (TAB-Seq) [57] on each of 20 samples (5 independent samples per condition from dorsal and ventral DG; 14.8-fold genome coverage per sample), a dataset unprecedented in its scale.

Each of the three forms of methylation exhibited a distinct genomic distribution in dorsal and ventral DG, leading to clear separation of dorsal and ventral samples in terms of methylation principal components (Appendix A.3A). A striking example is the locus containing *Nr2f2* (*COUP-TF2*), a developmental factor upregulated in ventral DG [40, 58]. The gene body of *Nr2f2* is surrounded by a large, 50 kbp DNA-methylation valley (DMV) that is dorsally hypomethylated in terms of mCG, mCH and hmCG (Fig. 2.2A, boxes i,ii,iv). The opposite pattern, ventral hypomethylation, prevails within the gene body of the shorter isoform, *Nr2f2.2* (box iii), consistent with the strong ventral-specific expression of this gene (>4-fold). The presence of large DMVs with both dorsal and ventral hypomethylation signatures at this locus illustrates the complex, region-specific relationship between DNA methylation and gene expression. We found additional DMVs associated with differentially expressed transcription factors such as *Nr2f1*, as

well as the developmental patterning factor *Pax7* (Appendix A.4).

Non-CG methylation (mCH) accumulates within neurons during post-mitotic maturation in the first 4 weeks of life in mouse frontal cortex [4] and accounts for 25% of methylcytosines in adult mouse DG [59]. Genome-wide, we found nearly twice as much mCH in ventral compared with dorsal DG ($p < 0.01$, Fig. 2.2B). This finding could be explained if increased neurogenesis in dorsal DG leads to a higher proportion of immature neurons, which may lack mCH. Global mCG and hmCG levels were equivalent in dorsal and ventral DG, and EE had no effect on global methylation levels. We did not detect significant hydroxymethylation at non-CG sites, consistent with cortical neurons and embryonic stem cells [4, 60].

A key advantage of whole-genome DNA methylation profiling is the ability to identify differentially methylated regions (DMRs), often far from any gene body, that mark tissue-specific gene regulatory elements [5, 47]. We found 23,000 DMRs that were hypomethylated in the dorsal relative to ventral DG [61] (hereafter called dorsal DMRs; mean methylation difference 26% $\pm$ 4.5% s.d.; Appendix A.3H, Supplementary Data), covering 4.45 Mbp or 0.16% of the genome in total (Fig. 2.2C). In contrast, we found only 587 DMRs hypomethylated in ventral relative to dorsal DG (hereafter called ventral DMRs), covering 84 kbp. This strong bias, with 40-fold more hypomethylated regions in the dorsal DG, contrasts with the balanced number of differentially expressed genes in dorsal and ventral DG (Fig. 2.2C,D), suggesting an asymmetric role for DNA methylation in region-specific gene regulation. Despite their small number, ventral hypomethylated DMRs marked key developmental patterning transcription factors (*Nf2f1/2*, *Pax3/7*), as well as *Efna5* and *Fgfr3* (Fig. 2.2C), which are linked to the proliferation, maintenance and survival of neural stem cells [62, 63].

2.2.4 DNA methylation correlates with repression at some genes

CG and non-CG DNA methylation are associated with reduced gene expression, while hmC associates with increased expression, as previously observed for frontal cortical neurons

[5, 4] (Appendix A.3B, C, F, G). We therefore examined whether dorsal-ventral differences in methylation correlated with region-specific expression. Genes upregulated in the dorsal DG were enriched for dorsal DMRs near the transcription start site (TSS) and throughout the gene body (Fig. 2.2D, green curve). These DMRs were also enriched at genes that are differentially expressed in EE compared to SH treated mice (Fig. 2.2E). Ventrally-upregulated genes showed a significant depletion of dorsal DMRs (Fig. 2.2D, purple curve) and an enrichment of ventral DMRs near the TSS (Appendix A.3D). Interestingly, dorsal DMRs were also enriched at genes that were up- and down-regulated in EE, although over half of dorsal up-regulated genes, and >98.5% of ventral up-regulated genes, contained no DMRs that could explain their region-specific differential expression (Fig. 2.2F, G, Appendix A.3D). These DMR-independent, differentially expressed (DE) genes included some with strong (>6-fold) regional specificity (e.g. *Grp*, *Cyp26b*, Appendix A.3E). DNA methylation may thus play a targeted role in controlling regional differentiation through key transcription factors. These factors could then sustain differential expression programs in a methylation-independent manner.

2.2.5 Impact of enrichment on DNA methylation in DG

EE enhanced the epigenetic distinction between dorsal and ventral DG, leading to detection of nearly 60% more dorsal DMRs in EE (16,156 DMRs) compared with SH-reared (10,185) animals (Fig. 2.2C). However, only a small number of regions were statistically significant DMRs when using the same criteria to directly compare SH and EE conditions (390 hypo-methylated, 595 hyper-methylated in EE). These DMRs did not overlap between the dorsal and ventral regions. We reasoned that EE-dependent changes in DNA methylation may be enriched within the relatively abundant dorsal DMRs, and thus focused our analysis on these sites. Upon averaging over all dorsal DMRs, we found lower dorsal DNA methylation levels in EE compared with SH at both CG ($p = 0.032$) and non-CG (CA, $p = 0.049$; CT, $p = 0.017$) sites (Fig. 2.3A, B, Appendix A.5). Ventral DNA methylation was not significantly different between EE and SH. Dorsal DMRs

were highly methylated in the fetal mouse cortex [4] and subsequently began losing methylation by one week of age (Fig. 2.3C). Dorsal DMRs thus mark regions that become demethylated during neuronal development. The decreased methylation of these regions in EE-reared mice is consistent with a higher proportion of immature neurons due to enhanced neurogenesis in the dorsal DG [54]. Further supporting this interpretation, we observed that most genes up-regulated by EE were also up-regulated in dorsal relative to ventral DG (Fig. 2.3D).

2.2.6 *NeuroD1* binding sites enriched at dorsal DMRs

To address the functional significance of DG DMRs, we analyzed the enrichment of transcription factor DNA sequence motifs [64] (Fig. 2.3A-D). Dorsal DMRs were strongly enriched for binding motifs of NeuroD1 ($p < 10^{-200}$, hypergeometric test), a basic helix-loop-helix transcription factor that is essential for maturation of newborn hippocampal neurons [65, 66, 67]. Dorsal DMRs were also enriched in motifs of the MEF2 family of transcription factors involved in neuronal differentiation [68] (Fig. 2.4A). By contrast, treatment-related DMRs hypomethylated in EE relative to SH were enriched for AP-1 family motifs, indicating activation of binding sites for the immediate early genes *Fos* and *Jun* (Fig. 2.4C). This is consistent with our transcriptomic data (Appendix A.2I,J) showing up-regulation of *Fos* and *Fosb* in EE treated mice, and implicates AP-1 signaling as a target for the effects of EE.

Treatment-related DMRs, including both those that are hypo- and hyper-methylated in EE, are enriched with binding motifs for *Grhl2* (Fig. 2.4C,D), a developmental factor that may contribute to survival of neuronal progenitors via its expression in non-neuronal cells [69]. Consistent with a potential glial role, *Grhl2* mRNA is expressed at a low level in our data from dentate gyrus (0.087 ± 0.3 TPM), as well as in data from dentate granule cells [40].

To validate the motif analysis, we examined DNA methylation in the dorsal DG at experimentally determined NeuroD1 binding sites from a previous study of in vitro neuronal differentiation [67]. We found a significant overlap of NeuroD1 ChIP-Seq peaks with dorsal

hypomethylated DMRs (67 peaks within 500 bp of a DMR; $p = 1.8x10^{-11}$, hypergeometric test; Supplementary Data). 48 genes contained NeuroD1 peaks collocated with a DMR (Fig. 2.4E). The vast majority of these genes, including *Tmem2* and *Epha8*, were significantly differentially expressed between dorsal and ventral DG (41/48); however, we also found NeuroD1 peaks overlapping DMRs in non-DE genes such as *Cog1* (Fig. 2.4E). Consistent with the motif enrichment analysis, we found lower mCG in dorsal compared with ventral DG at NeuroD1 ChIP-Seq peaks (Fig. 2.4F). Although we found no effect of EE on mCG levels at these sites, there was a significant reduction in mCA at these sites specifically in the dorsal, but not ventral, DG (p=0.0006, Fig. 2.4G). The EE-associated differences in mCA were highly localized to the NeuroD1 binding site (Fig. 2.4H). Thus, subregion-specific, environmental influences on dentate gyrus appear to reflect dynamic epigenetic modifications at non-CG sites within NeuroD1 transcription factor binding regions that are linked to neuroplasticity, including neurogenesis.

2.3 Discussion

Our study integrates whole-genome, base-resolution DNA methylation and hydroxymethylation data with gene expression (RNA-Seq), in vivo structural MRI and immunohistochemistry, in a mouse model of peripubertal environmental enrichment. Environmental enrichment is a form of early experience that stably alters neural development and behavior in rodent models [70]. Using these multi-modal datasets we have identified subregion-specific transcriptomic and epigenomic influences of enriched experience in the dorsal and ventral DG. We find that the magnitude of the molecular differentiation of the dorsal and ventral hippocampus is influenced by early experience. Based on our data and analysis, we can begin to propose a unified model of epigenomic and transcriptional regulation in the DG integrating both region-specific and environmental enrichment effects (Fig. 2.5).

Lesion studies and connectivity profiles of the hippocampus have suggested that the

dorsal hippocampus is critical for spatial cognition, whereas the ventral region is associated with emotional processing and stress responses [24]. There are substantial expression differences along the dorsal-ventral axis of the DG, as well as hippocampal subregions CA1, CA2 and CA3 [24, 40, 34]. However, regulatory mechanisms that could support these differences remain unexamined. Our data bridge this gap, linking dorsal-ventral DNA methylation differences with transcription. For example, we identified hypomethylated regions in the ventral DG at *Pax3* and *Pax7*. These transcription factor genes restrict ventral fate in the spinal cord and could play a similar role in the hippocampus [71]. These results extend our knowledge of the substantial epigenomic and transcriptional differences that parallel the functional specialization of the dorsal and ventral DG [24, 40, 51, 72].

The high level of correlation ($r = 0.988$) among transcriptomes from our five independent samples allowed us to detect 3,497 differentially expressed genes with high statistical confidence, far more than were previously reported in purified granule cells [40]. This illustrates that gene expression profiling in intact tissues is a valuable complement to cell type specific approaches, which may perturb the cellular transcriptome in the course of cell type purification. While the transcriptional differences we observe between dorsal and ventral DG are substantial, they are of a smaller magnitude than differences among cortical cell types (Appendix A.2B) [5]. For example, there are 4.7-fold more DE genes (using a cutoff $>$2-fold differential expression) when comparing cortical excitatory neurons with PV- or VIP-positive fast-spiking interneurons.

In contrast with the widespread differential gene expression between dorsal and ventral DG, we found a more limited number of DNA methylation and hydroxymethylation differences (mCG and hmCG). We did find a 2-fold higher abundance of mCH throughout the genome in the ventral compared with dorsal DG. While notable DNA methylation differences at key transcription factor and ventral patterning genes were negatively correlated with gene expression, overall our data suggest that many dorsal-ventral gene expression differences cannot be directly linked to DNA methylation differences.

Adult neurogenesis in the DG is enhanced by EE [70], but the molecular mechanisms mediating this process remain unknown. Brain-derived neurotrophic factor (BDNF) is upregulated at the mRNA level in mouse hippocampus following 3-4 weeks of exposure to EE [73], while EE-induced adult neurogenesis was blocked in a heterozygous knockout (*Bdnf*+/-) [74]. Similarly, mRNA for vascular endothelial growth factor (VEGF) is upregulated in hippocampus upon exposure to EE, and manipulations that increase or decrease VEGF levels cause corresponding increases and decreases in neurogenesis [75]. We did not detect differential expression of *Bdnf* or *Vegf* in the dorsal or ventral DG, suggesting these factors may be upregulated in other hippocampal regions. We did identify up-regulation in EE of mRNA for dopamine receptor D1 (*Drd1*), which is expressed in dentate granule cells [76] and gates long-term changes in synaptic strength [77, 78], and the opioid neuropeptide *Penk* that is expressed in a subpopulation of DG granule cells [51]. We also found activation of immediate early genes (IEGs), consistent with increased synaptic activity. Exposure to a novel environment activates IEG transcription in DG granule cells that can be detected by single nuclei sorting followed by RNA-Seq [34]. Our data suggest IEGs are also activated by long-term exposure to an enriched environment, which includes continuous introduction of novel objects as well as social and physical stimulation. Importantly, by performing 5-fold replicate experiments on independent biological samples, each drawn from 10-12 animals, we could stringently assess the reproducibility and robustness of gene expression changes.

Changes in DNA methylation can mediate long-lasting environmental effects on gene expression and behavior [41]. EE induces stable behavioral changes [21], yet the role of DNA methylation has not been examined. In our EE cohort, we observed a 31% upregulation of *Gadd45b*, involved in activity-induced DNA demethylation [55]. We found few DMRs in a direct comparison of EE and SH raised animals, indicating that individual DNA methylation changes in this paradigm may fall below the detection threshold for whole genome bisulfite sequencing. We did observe an effect of EE in modulating DNA methylation at dorsal-ventral

DMRs. There were 59% more dorsal DMRs (methylation significantly lower in dorsal compared to ventral DG) in our EE cohort compared to SH. These DMRs were enriched for binding sites of the neurodevelopmental transcription factor, NeuroD1, which is upregulated in maturing adult newborn neurons. These genomic regions also showed significantly lower methylation in EE compared to SH at CA and CT dinucleotides, suggesting an effect of early experience on a largely brain-specific form of methylation. These findings could be explained by changes in methylation within existing cells, changes in the proportion of maturing newborn neurons, or a combination of both. We also examined the role of 5-hydroxymethylcytosine (5hmC) in EE. Ten-eleven translocation (TET) family of enzymes can catalyze the conversion of 5-mC to 5-hmC [79]. Although its function is not fully understood, 5-hmC may represent an intermediate state produced during demethylation. We found 5-hmc was positively correlated with transcription, supporting the idea that 5-hmC mediates transcription.

Previous work suggests a functional distinction between the dorsal and ventral DG, and our work shows the two poles are differently affected by EE [24]. We detected 80 more differentially expressed genes in the dorsal than the ventral DG in response to EE. In addition, as noted above, the EE-reared animals showed many more dorsal DMRs (16,156) compared to SH treated animals (10,185). These regional differences may be consistent with a greater enhancement of neurogenesis by EE in the dorsal as compared to ventral DG [80].

Although our data are unprecedented in resolution and sample size, there are still some challenges to identifying the source of transcriptional and methylation changes in tissue from a heterogeneous and dynamic cell population like the DG. For example, we cannot distinguish between changes in DNA methylation occurring in a stable population of mature neurons, and changes to the proportion of immature and newborn neurons due to increased neurogenesis. Neurons in all stages of the maturation process coexist within the adult DG, and our data represent a mixture of signals from stem cells and immature and mature neurons. Similarly, the dorsal-ventral differences in DNA methylation could be driven by differences in cell type composition

between the two regions, or a discrete or graded difference between the DG neurons in the dorsal and ventral poles. Here we attempted to better understand the heterogeneity of our tissue by correlating our RNA-seq data with known neuron type transcriptional profiles (Appendix A.2P) [52]. Although the strongest correlation between our dDG and vDG bulk tissues was with neurons ($r = .89$), we also found substantial correlations with gene expression patterns in other cell types. Thus, it remains difficult to determine to what extent regional differences and EE-induced changes in cellular heterogeneity may account for our results. Future studies, including single cell assays, could address these limitations and better characterize transcription and DNA methylation in maturing newborn neurons and adult DG neurons [51, 34, 81, 82].

Overall, our transcriptome and DNA methylation data support a model of regional and environmental effects on the molecular profile of DG neurons (Fig. 2.5). First, assuming only mature neurons have mCH [4] and that the mCH levels in mature dorsal and ventral dentate granule cells are similar, our finding of lower mCH in dorsal DG suggests a higher proportion of immature neurons in this region. Second, regional differences in expression of RGL and NSC markers suggest an increased proportion of NSCs in dDG and increased RGLs in vDG. This distinction is further supported by the preponderance of dorsal DMRs over ventral DMRs and their enrichment for the binding of the neuronal differentiation factor, NeuroD1. Finally, by promoting neurogenesis in the dDG, EE has the effect of further increasing the proportion of immature neurons in this region, leading to low mCG and mCA levels at dorsal DMRs and NeuroD1 binding sites.

2.4 Methods

2.4.1 Animals and environmental enrichment

All procedures were performed in accordance with the guidelines established by the Canadian Council on Animal Care (CCAC) with protocols approved by the McGill University

Facility Animal Care Committee (FACC). Male C57/Bl/6 mice were bred at the Douglas Institute to avoid transportation stress. Mice were weaned on postnatal day 22, and siblings were assigned to either standard or enriched housing conditions. Standard housed animals were raised in groups of three male mice from different mothers in a 30 x 18 cm cage. The enriched group contained 12 male mice, housed in a larger rectangular plexiglass cage (78 x 86 cm) with a plexiglass top, which contained a variety of toys such as running wheels, a bridge, and novel objects. Toys were changed weekly. For animals in both conditions, food and water were provided ad libitum, and bedding was changed biweekly, cleaning the cages with a Peroxyguard solution. Animals remained in the respective housing conditions for eight weeks. Mice were sacrificed on age day 80 (post sexual maturation) between 10:30 am to 12 pm. A cohort of 10 mice per housing condition was used for magnetic resonance imaging (MRI). A separate cohort was used for sequencing assays with five samples per housing condition, and each sample was composed of tissue from 10-12 mice. A separate cohort of male mice ($n = 20$) was used for hippocampal neurogenesis study.

2.4.2 Tissue collection for MRI

Mice were perfusion-fixed on postnatal day 80 as previously described [83]. Briefly, mice were perfused via the left ventricle using 30 ml of room-temperature (25°C) phosphate-buffered saline (PBS) (pH 7.4), 2 mM ProHance (gadoteridol, Bracco Diagnostics Inc., Princeton, NJ), and 1 μl/ml heparin (1000 USP units/ml, Sandoz Canada Inc., Boucherville, QC) at a rate of approximately 1 ml/minute. Next, 30 ml of 4% paraformaldehyde (PFA) in PBS containing 2 mM ProHance was infused at the same rate. After fixation, the heads, skin, ears, and lower jaw were removed and the skull was allowed to postfix in 4% PFA at 4°C for 24h. The samples were then placed in a solution of PBS, 2 mM ProHance, and 0.02% sodium azide (sodium trinitride, Fisher Scientific, Nepean, ON) and stored at 4°C until imaging.

2.4.3 Magnetic resonance imaging and analysis

Anatomical whole-brain images were acquired 16 at a time using a multi-channel 7.0-T scanner and custom-built 16-coil solenoid array (Varian Inc., Palo Alto, CA) [84, 85]. Brains were imaged using a T2-weighted, 3D fast spin-echo sequence at 56-micron isotropic resolution (MRI parameters: TR = 2000 ms, echo train length = 6, TEeff = 42 ms, field-of-view (FOV) = 25 x 28 x 14 mm^3 and matrix size = 450 x 504 x 250, imaging time = 11.7 h). To correct for small geometric distortions resulting from imaging in coils not in the centre of the magnetic field, coil-specific MR images of precision-machined phantoms were registered to a computed tomography (CT) scan of the same phantom. The resulting distortion correcting transformations were then applied to all acquired images in a coil-specific manner.

To determine the effect of housing condition on brain anatomy, all images in the study were aligned using an automated image registration pipeline as described previously [84, 86]. All registrations were performed with a combination of mni_autoreg tools [87, **?**] and Advanced Normalization Tools (ANTS) [88]. Briefly, the images were first linearly aligned using a series of global rotations, translations, scales, and shears. They were then locally aligned via an iterative nonlinear process which brings all images into precise anatomical alignment in an unbiased fashion [86, 89]. The output of this automated registration process is a study-specific consensus average, representing the average anatomy of all mice in the study, along with deformation fields that encode how each individual image differs from the study average [84, 86]. After registration, a manually labeled MRI atlas delineating dorsal and ventral hippocampus was aligned to the study average. This was used in combination with the deformation fields to calculate the volume of the dorsal and ventral hippocampus for each subject in the study in an automated and unbiased fashion [84, 86]. The effect of housing condition on dorsal and ventral hippocampal volume was assessed using Student's *t*-tests. The interaction effect between housing condition and region on volume was assessed using a linear mixed effects model with random intercepts for each mouse using the lmerTest package [90]. Image analysis was performed using the R statistical language (R

Core Team, 2016, https://www.R-project.org) and the RMINC library (https://github.com/Mouse-Imaging-Centre/RMINC). Error bars represent 95% confidence intervals.

2.4.4 Immunohistochemistry

Animals were intraperitoneally injected with BrdU (100 mg/kg, 20 mg/ml, Cat# B5002, Sigma-aldrich) twice on 2 consecutive days at postnatal day 80. 30 days following the last injection, the animals were sacrificed via transcardial paraformaldehyde (4% in 1x phosphate-buffered saline) perfusion. The sliced brain sections were processed for immunohistochemistry using Anti-BrdU antibody (Abcam, Cat# ab6326, 1:400) and visualized with DAB (Cat# SK-4100, Vector Laboratories). BrdU immunoreactive cells were counted in the subgranular zone and granule cell layer region in dorsal (8-12 section, 80 μm apart, bregma -1.34 to bregma -2.30) and ventral (8-10 sections, Bregma -2.92 to Bregma -3.64) [91] hippocampus per animal under VS120 virtual slide microscope (Olympus). The number of labeled cells per dentate gyrus was statistically tested using a two-way analysis of variance (ANOVA) with housing condition and marginal region as main effects.

2.4.5 Tissue collection for sequencing assays

Tissue collection consisted of rapid removal of the brain, followed by flash freezing and storage at -80°C. Frozen brains were sliced coronally at 200 μm thickness until reaching bregma -2.30. The brains were then removed from mounting position, rotated, and remounted to the mounting position for horizontal slicing ventral dentate tissue. Horizontal sections were sliced from interaural 3.24 mm to 0.92 mm [91]. A 300 μm diameter puncher was used to punch dorsal and ventral dentate gyrus region separately.

2.4.6 RNA and DNA extraction

RNA and DNA extraction were performed from the same sample using Qiagen Allprep DNA/RNA Mini kit (Qiagen, Cat# 80204.). We performed on-column DNase I treatment during RNA extraction and on-column RNaseA treatment during DNA extraction. RNA was examined by Bioanalyzer 2100 (Agilent technologies, Santa clara, USA).

2.4.7 RNA-Seq collection

The RNA libraries were prepared in McGill University and Genome Quebec Innovation Centre using Illumina TruSeq Stranded total RNA LT set (Cat# RS-122-2301, Illumina Canada Ulc.). Paired-end, 100bp read-length RNA-seq was collected using HiSeq 2000 at a depth of 30 M sequencing.

2.4.8 Validation of RNA-Seq results by digital Nanostring

Housing differences in RNA-seq were validated with Nanostring on 48 randomly selected differentially expressed genes. 100ng of tissue were sent to Jewish General Hospital (Montreal, Quebec, Canada) for expression quantification using NanoString nCounter XT-GX (NanoString Technologies, Inc., Seattle, WA, USA). Probes were designed to hit the maximum number of validated transcript variants while minimizing the cross-reactivity of the probes. Scanned data was normalized using Nanostring-provided housekeeping genes and analyzed using nSolver Analysis Software 2.6 (NanoString Technologies, Inc., Seattle, WA, USA). Comparison of mRNA fold change between RNA-seq and NanoString shows consistent results (Appendix A.2, K-N)

2.4.9 TAB-Seq and MethylC-Seq

DNA from the same samples was separated for TAB-seq and MethylC-seq library preparation. TAB-seq measures levels of 5 hydroxymethylation (5-hmC). Protection and oxidation

portions of library preparation were performed in-house using the Wisegene kit as described in Yu et al [57]. Three spike-in control DNAs, lambda DNA (Cat# D1501, Promega), 5mC (Cat# S001, Wisegene) and 5hmC (Cat# S002, Wisegene) were added to each sample (2.5 μg of total DNA) before DNA shearing, in order to evaluate the bisulfite conversion efficiency, the protection rate of 5 hmC, and the oxidation rate of TET. In 5hmC control spike-in DNA, due to the impurity of commercial 5hmdCTP and slow oxidation of 5hmC upon exposure to air, the actual abundance of 5hmC at each cytosine site is not 100% hydroxymethylated. Therefore, we ran the same batch of 5hmC spike-in control in another bisulfite sequencing to examine its real 5hmC abundance.

Bisulfite conversion was then performed at the Genome Quebec Innovation Centre on the processed TAB-seq sample, as well as 1 μg of DNA for the MethylC-seq library. Methylated and unmethylated DNA sets (Cat# D5017, pUC19 DNA set, Zymo research) were added as spike-in controls (2 ng spike-in control in 1 ug DNA) to evaluate bisulfite conversion efficiency. The whole genome bisulfite sequencing (WGBS) libraries were prepared using NimbleGen SeqCap Epi Enrichment System (Cat# 07145519001, Roche NimbleGen, Inc.). Library amplification was done using KAPA HiFi Hotstart Uracil + DNA polymerase (Cat# KK2802, Kapa Biosystems).

2.4.10 Data analysis

All analyses were conducted in either Matlab or Python with packages including Numpy, Scipy, Pandas, Matplotlib and Sklearn. All data were aligned to the mm10 (GRCm38) reference genome, and genes were defined using Gencode annotation vM7 level 3 transcriptome (http://www.gencodegenes.org/). Browser representations were created using AnnoJ (http://www.annoj.org) [10]. Pearson correlations were used except where stated otherwise. P-values were ¡0.01 unless otherwise stated.

2.4.11 Differential expression

RNA-Seq data were aligned using STAR Aligner in quantMode to obtain gene counts [92]. Differentially expressed genes were identified using generalized linear models and contrasts in EdgeR [93]. We only retained genes with counts > 10 in at least two samples for the analysis. In addition, we excluded one SH sample due to high expression of the long noncoding RNA, *Xist*, which is only expressed in females. We then tested the below null hypotheses to identify differentially expressed genes by region (1) and treatment in the dorsal (2) and ventral (3) dentate gyrus. Benjamini Hochberg was used to control the false discovery rate ($q < .05$).

$$DorsalSH - VentralSH = DorsalEE - VentralEE$$

$$DorsalEE - DorsalSH = 0$$

$$VentralEE - VentralSH = 0$$

2.4.12 Differential methylation analysis

Whole genome bisulfite sequencing data were mapped using Methlypy [4]. The non-conversion rate (NCR) was estimated using a fully unmethylated phage lambda DNA spike-in. NCR was found to be low across all samples ($.43\% \pm .021\%$). Methylation values were corrected for the NCR using the following maximum likelihood formula, where m is the number of methylated base calls and c is the total number of base calls:

$$mC = g\left[\frac{m/c - NCR}{1 - NCR}\right]$$

$$g[x] = max[x, 0]$$

Differentially methylated regions ($\geq$15% methylation difference, $p < .05$) at CG dinucleotides were identified using DSS [61, 94]. To examine the link between differential expression and DMRs, we computed an enrichment score (the density of DMRs per gene per 1 MB) as a function of distance from the transcription start site (TSS). Enrichment scores were compared between differentially expressed and non-differentially expressed genes using a hypergeometric test.

Tet-assisted bisulfite sequencing (TAB-Seq) is a methodology for measuring genome-wide 5-hydroxymethylation that consists of three main steps: protection, the binding of a glycosyl group to hydroxymethylated cytosines; Tet oxidation, the demethylation of non-glycosylated methylated cytosines; and bisulfite treatment, conversion of all unmethylated cytosines to uracils [57]. Upon sequencing, only 5-hydroxymethylated cytosines should still be cytosines. To measure the inefficiency of each of these steps, a fully hydroxymethylated (pUC19) and a fully methylated (lambda phage) spike in are included. Corrected hydroxymethylation levels were computed using the below formula with variables r_{TAB}(bisulfite non-conversion in the TAB-Seq data, estimated via Lambda DNA in the CH context), s_{TAB}(non-oxidation in the TAB-Seq data, estimated using Lambda DNA in the CG context), t_{TAB}(non-protection in the TAB-Seq data, estimated using pUC19), and p_{mC} (the fraction of mC+hmC):

$$p_{hmC} = g\left[\frac{q_{TAB} - s_{TAB}p_{mC} - r_{TAB}(1-p_{mC})}{1-t_{TAB}}\right]$$

$$= g\left[\frac{q_{TAB} - r_{TAB} - (s_{TAB} - r_{TAB})p_{mC})}{1-t_{TAB}}\right]$$

Finally, we examined DMRs for enrichment of transcription factor binding sequence motifs using Homer [64]. For this analysis, sequences within 200 bp of each DMR center were included. We examined the overlap of DMRs with ChIP-Seq data [67].

2.4.13 Gene ontology analysis

Functional examination of gene sets was performed via gene ontology analyses include EnrichR (http://amp.pharm.mssm.edu/Enrichr/) and Metacore 6.27 (build 68571).

2.4.14 Data availability

Raw and processed data reported in this study are available via the Gene Expression Omnibus with accession GSE95740, https://www.ncbi.nlm.nih.gov/geo/. A browser visualization of genomic data is at http://brainome.ucsd.edu/mouse_dentategyrus.

2.5 Acknowledgements

Chapter 2, in full, is a reprint of the material as it appears in *Nature Communications* 2018. Tie-Yuan Zhang*, Christopher L Keown*, Xianglan Wen, Junhao Li, Dulcie A. Vousden, Christoph Anacker, Urvashi Bhattacharyya, Richard Ryan, Josie Diorio, Nicholas O'Toole, Jason P. Lerch, Eran A. Mukamel, Michael J. Meaney. The dissertation author was the co-investigator and co-first author of this paper.

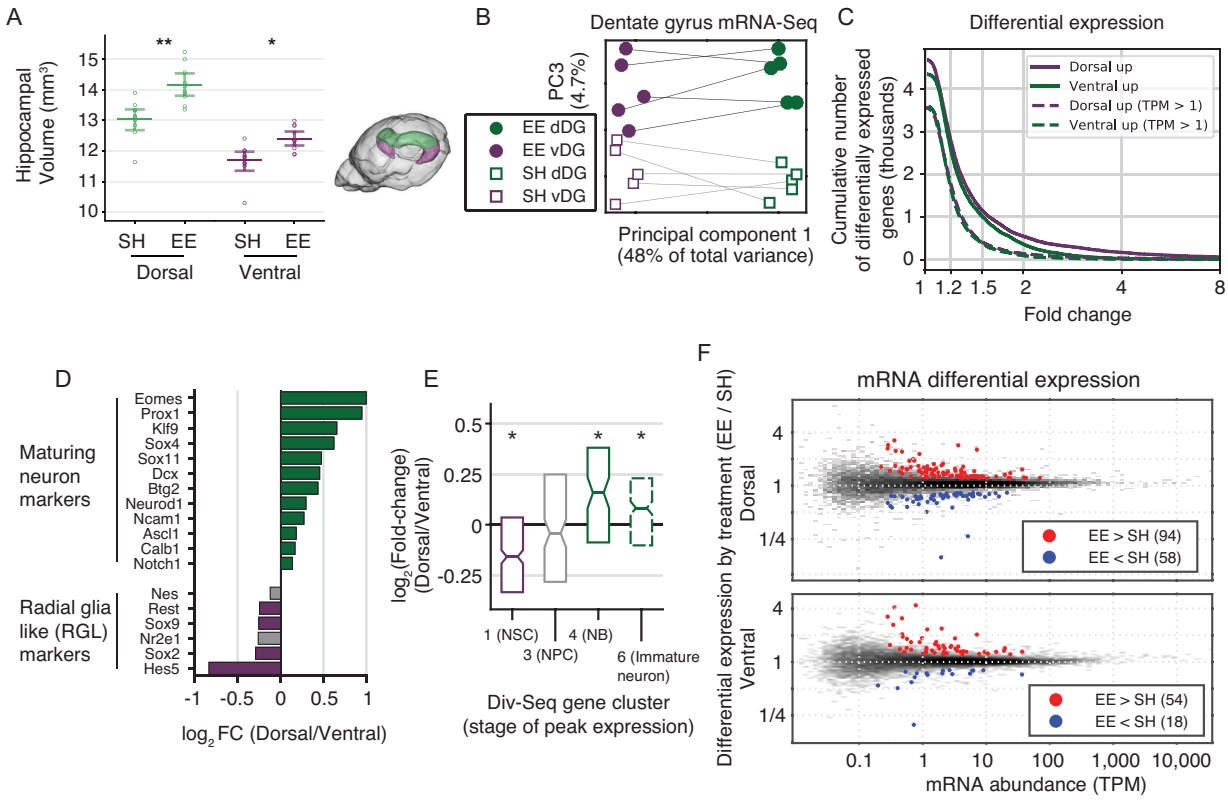

Figure 2.1: Transcriptional effects of enriched environment (EE) are greater in dorsal than ventral dentate gyrus (DG). (**A**) High-field structural MRI shows enlarged hippocampus in EE-treated animals. (**B**) DG transcriptome principal components separate dorsal and ventral samples (PC1), as well as standard housing (SH) vs. EE reared animals (PC3). Dorsal and ventral samples from the same mice are connected by lines. (**C**) The cumulative number of genes differentially expressed in dorsal vs. ventral DG (FDR < 0.05) as a function of the minimum expression difference cutoff. Here we consider all genes with > 10 RNA-Seq read counts in ≥ 2 samples (solid lines), or with TPM > 1 in ≥ 3 samples (dashed). (**D**) Maturing neuron and radial glia like (RGL) markers [42, 53] are enriched in dorsal and ventral DG, respectively (gray bars: not significant). (**E**) Clusters of genes active in RGL or immature neurons in Div-Seq data [51] are enriched in dorsal and ventral DG, respectively. NSC: neuronal stem cell; NPC: neural progenitor; NB: neuroblast. (**F**) Twice as many genes are differentially expressed in EE vs. SH in dorsal compared with ventral DG.

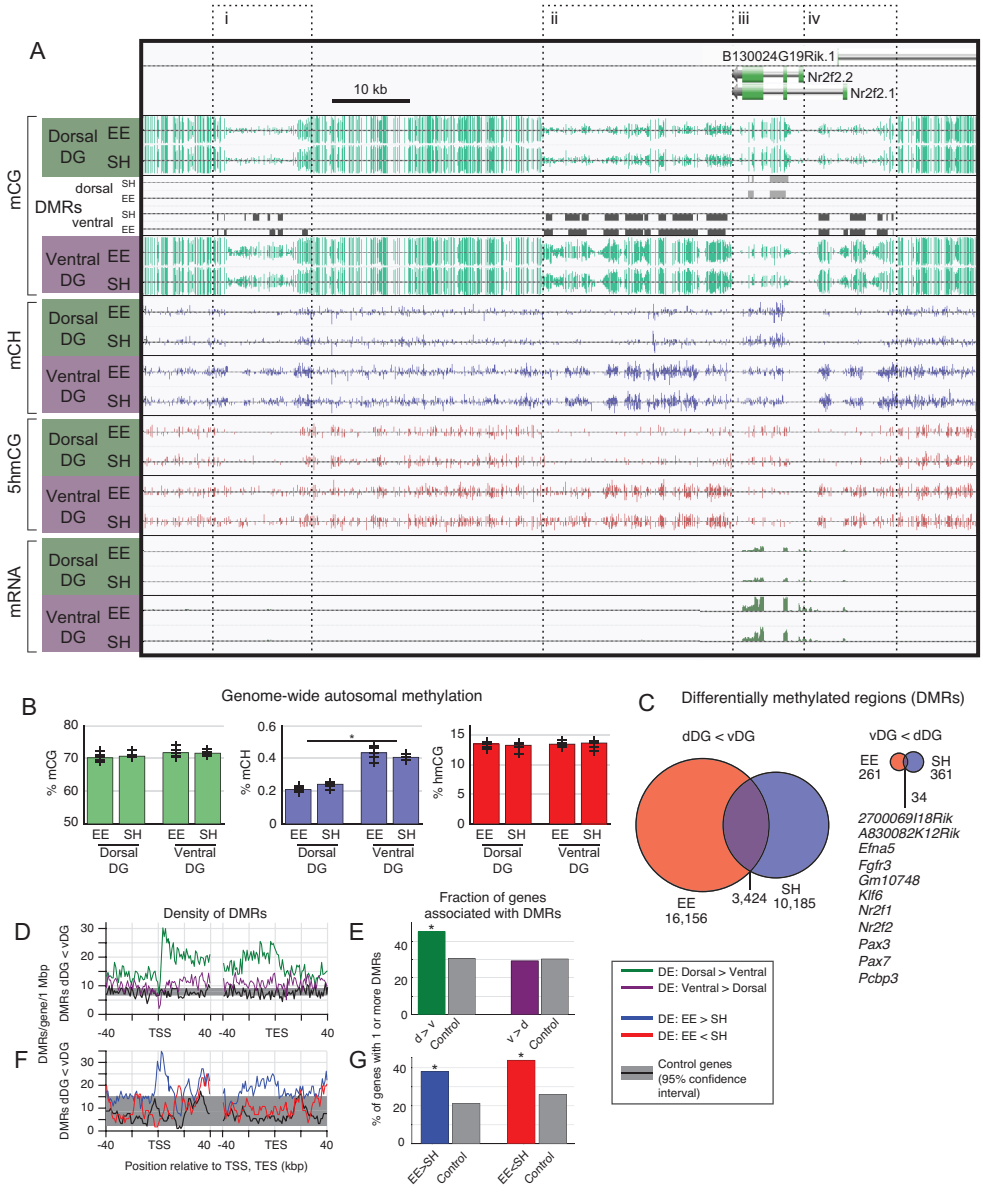

Figure 2.2: Reduced DNA methylation in dorsal dentate gyrus associated with expression differences. (**A**) Browser view of the locus containing development factor *Nr2f2* (*Coup-TF2*) shows bidirectional differentially methylated regions (DMRs) and corresponding differences in DNA methylation (mCG, mCH), hydroxymethylation (hmCG), and mRNA expression. (**B**) The genome-wide mCH level is 50% lower in dorsal compared with ventral DG; mCG and 5hmCG did not differ (+ symbols indicate levels for individual samples). (**C**) The vast majority of region-specific DMRs are hypomethylated in dorsal (dDG<vDG). The smaller number of ventral hypomethylated DMRs (vDG<dDG) includes many key developmental transcription factors. (**D**,**E**) DMRs are enriched at differentially expressed (DE) genes. Gray shaded region: 95% confidence interval from control genes with equivalent mean expression. (**F**,**G**) Over half of DE genes contain no DMR.

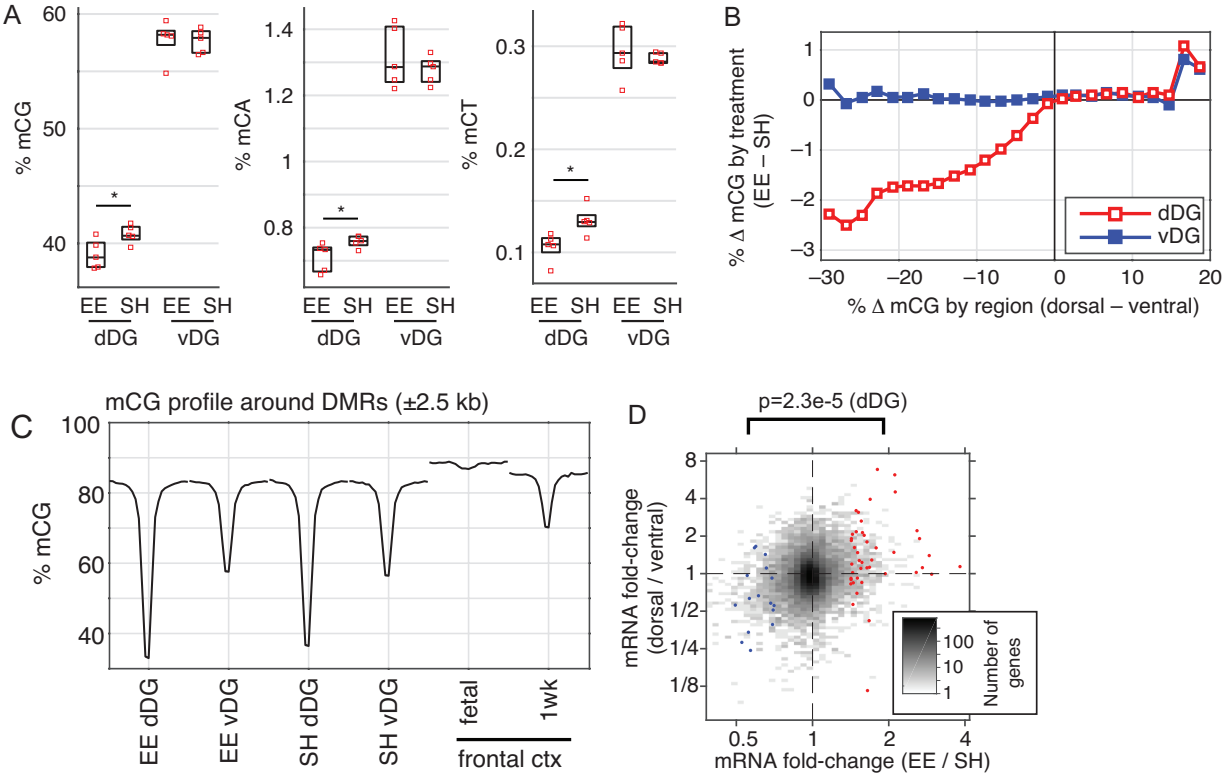

Figure 2.3: Greater dorsal-ventral differentiation of DNA methylation in enriched environment. (**A**) At DMRs hypomethylated in dDG, dorsal DNA methylation is lower in EE compared with SH reared animals at CG, CA and CT sites ($p < 0.05$, ANOVA). Ventral methylation is unaffected. (**B**) Median difference in mCG between EE and SH samples across all genomic bins (1kbp) stratified by regional (dorsal-ventral) difference in mCG shows a strong effect in the dorsal, but not ventral, DG. (**C**) Mean mCG profile centered on dorsal DMRs in DG, as well as in fetal and 1 week old frontal cortex [4]. (**D**) Genes that are up-regulated in EE are also enriched in dDG.

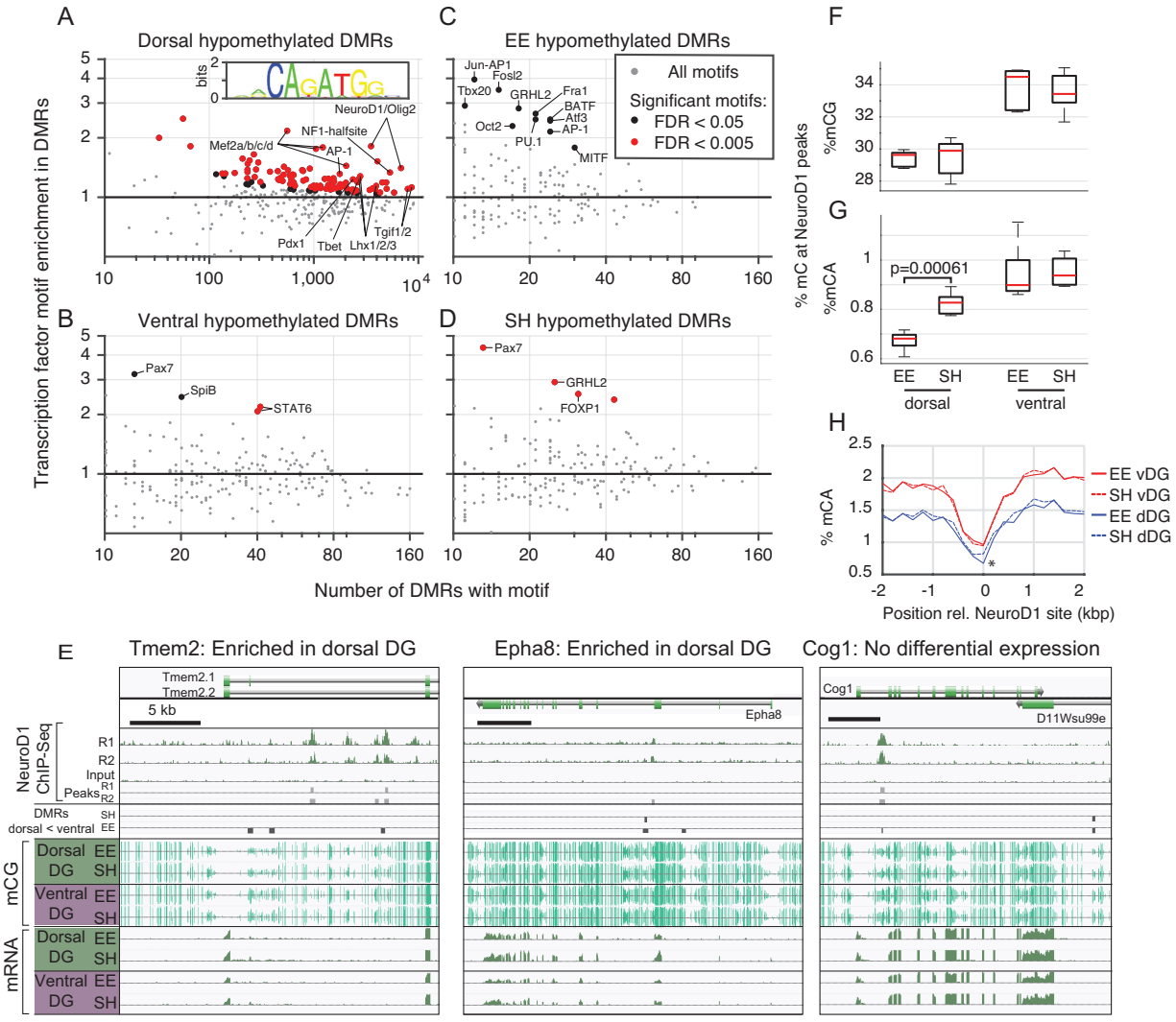

Figure 2.4: Transcription factor binding sites are enriched at DMRs. (**A**-**D**) Known transcription factor binding site sequence motifs are significantly enriched within DMRs. (**A**) Dorsal DMRs are enriched for motifs of developmental and neuronal differentiation TFs, including NeuroD1. Inset: sequence logo of *de novo* sequence motif matching the NeuroD1 binding motif. (**C**) EE DMRs are enriched for binding sites of AP-1 family immediate early genes. (**C**-**D**) EE and SH DMRs are enriched for GRHL2 motifs. (**E**) Dorsal DMRs significantly colocalize with experimentally determined binding sites of NeuroD1 [67] at dorsally enriched genes (*Tmem2*, *Epha8*) and at some genes with no significant differential expression (*Cog1*). (**F**-**H**) mCA is significantly reduced in dDG at NeuroD1 binding sites.

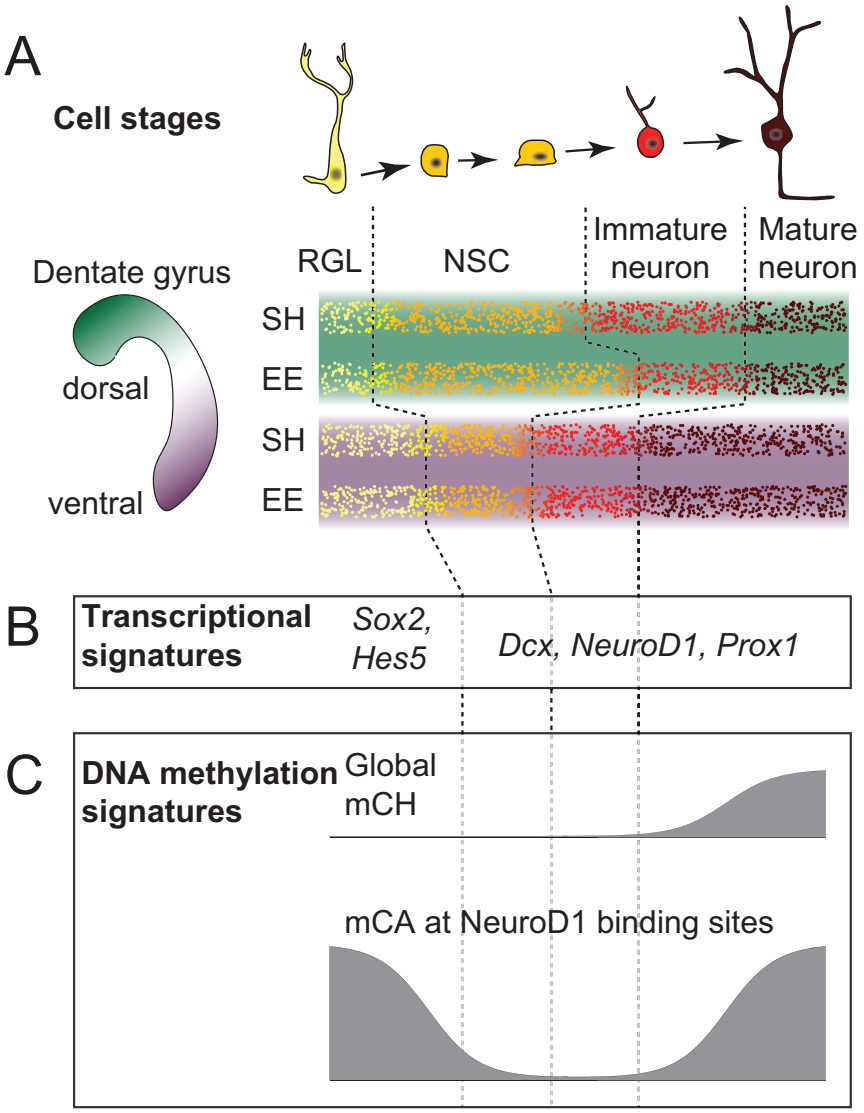

Figure 2.5: A model for epigenetic regulation of dorsal and ventral DG. (**A**): The cell stages occurring within the subgranular zone of the dentate gyrus are shown together with a schematic illustration of possible relative proportions consistent with our data. RGL: Radial glia-like progenitor; NSC: Neural stem cell. (**B**) Key genes associated with the RGL stage are up-regulated in ventral DG relative to dorsal DG. (**C**) We propose that mCH accumulates mainly in mature neurons.

Chapter 3

Allele-specific non-CG DNA methylation marks domains of active chromatin in female mouse brain

3.1 Introduction

In diploid mammals, the equivalence of the two parental alleles is violated by allele-specific epigenetic regulation in a small, but critical, subset of the genome. Genomic imprinting, or parent-of-origin-dependent gene regulation [95], is critical for embryonic development and plays a role in neuronal differentiation [96]. In females, epigenetic inactivation of one X chromosome silences transcription of most genes to equalize gene expression with males [25]. Both imprinting and X-chromosome inactivation (XCI) are critical to healthy brain development [28, 97]. Despite the importance of allele-specific gene regulation in the brain, the epigenetic mechanisms controlling these patterns are not completely known, in part, due to the challenge of allele-specific epigenomic profiling. In particular, DNA methylation patterns can reflect allelic asymmetries in autosomal gene regulation [9], but their correlation with XCI has not been fully

addressed.

XCI has unique advantages as a case study for the investigation of allele-specific epigenomic regulation. The inactivated allele is selected stochastically during early development and maintained through subsequent cell divisions [98], yielding a mosaic pattern of allelic expression in adult female tissues. Despite extensive inactivation of one X chromosome, some genes escape silencing and are expressed from the inactive X chromosome (Xi): $\sim 3\%$ of X-linked genes in mice [99] and 15% in humans [27]. Analysis of peripheral blood showed that XCI and escape from XCI are correlated with high or low levels of DNA methylation at CG dinucleotides (mCG) in promoter regions, respectively [29]. However, different epigenetic profiles may be associated with XCI and escape from XCI in the brain because the DNA methylation landscape of neurons is distinct from other cell types. In particular, neurons accumulate methylation at millions of genomic cytosines in CA and CT dinucleotides during postnatal brain development beginning at 1 wk of age in mice [9, 4]. This non-CG methylation correlates with reduced gene expression and inactivation of distal regulatory elements in a highly cell type-specific manner [5]. Although the functional relevance of non-CG methylation (mCH) is unclear, it is bound by the transcriptional repressor methyl-CpG binding protein 2 (MeCP2) as neurons mature, and is enriched at genes that are up-regulated in Rett syndrome [100, 101].

Mosaic XCI prevents discrimination of methylation on the active X chromosome (Xa) and Xi alleles by conventional methylome profiling. We reasoned that understanding the allele-specific distribution of neuronal mCH in the context of X inactivation and imprinting could yield new insights into this unique aspect of the brain epigenome. Therefore, we profiled allele-specific DNA methylation, as well as transcription, in mouse frontal cortex using a *Xist* mouse mutant hybrid in which the paternal allele was deterministically inactivated in all cells [99]. To assign sequencing reads to alleles, we used female F1 mice from crosses between C57BL/6 *Xist* mutant and *Mus spretus* wild-type mice [99], and analyzed species-specific genetic variants (~ 42 million single-nucleotide polymorphisms (SNPs), including 1.95 million SNPs on the X chromosome).

Our data reveal distinct allele-specific patterns of mCG and mCH at X-linked genes that reflect the accessibility of chromosomal domains during brain development. At autosomal imprinted regions, we found large domains of differential methylation that show a dissociation between mCG and mCH and point to independent regulation of these features of the neuronal epigenome.

3.2 Results

3.2.1 Allele-Specific Global Levels of CG and Non-CG DNA Methylation on Female X Chromosomes

Using female interspecific F1 transgenic mice with deterministic X inactivation [99, 30], we examined global levels of mCG and mCH on Xa and Xi in the adult frontal cortex. mCG is present throughout nonpromoter regions of the genome from the earliest stages of brain development, whereas mCH accumulates gradually during postnatal development starting at 1 wk of age in mice [4]. We therefore reasoned that the silencing of Xi, established during random XCI in the embryonic inner cell mass [98], may block the accumulation of mCH on Xi. By contrast, nonpromoter mCG may be less affected by the chromosomal inactivation because it is established and epigenetically inherited from the early embryonic stage [102].

Consistent with this reasoning, we found significantly increased levels of mCG at promoters on Xi (69.4%) compared with Xa (40.4%; P = 0.003, paired t test) and autosomes (30-33%; $p = 0.006$) (Fig. 3.1A). By contrast, mCG at nonpromoter regions was lower on Xi (75.6%) compared with Xa (85.6%; $p = 0.013$) and autosomes (84-85%; $p = 0.012$). Strikingly, Xi is nearly devoid of mCH (0.02%) compared with Xa and autosomes (1.01%; $p < 0.001$). Xi thus resembles mCH deserts: large regions (median size of 1.8 Mbp) on autosomes that lack mCH, are transcriptionally silent, and are marked by inaccessible chromatin [4]. The pattern of methylation is likewise less correlated across the two X alleles (mCG: $r = 0.27$, mCH: $r = -0.08$; correlation

using 10-kb bins) compared with the autosomes (mCG: $r = 0.90$, mCH: $r = 0.91$). These results suggest that Xi is largely inaccessible to the de novo DNA methyltransferase Dnmt3a, which is responsible for establishing mCH in neuronal genomes starting in the second postnatal week [101].

3.2.2 Differential Methylation Patterns at Genes Known to Escape XCI

A subset of X-linked genes escapes from XCI, allowing expression from Xi. Comparison of male and female brain samples from both mice and humans showed a striking enrichment of mCH in females within the gene bodies of several escape genes [4, 103]. Our allele-specific analyses show that this mCH signature of escape genes derives exclusively from the Xi. The pattern is exemplified by the allele-specific expression and methylation profiles of a known escape gene, *Kdm5c*, and two neighboring nonescape genes, *Iqsec2* and *Kantr* (3.1B). By sorting reads based on the presence of SNPs that vary between C57 and *spretus* genomes (Methods), we identified sequencing reads originating from Xa and Xi for both expression and methylation data. As expected, *Iqsec2* and *Kantr* were monoallelically expressed from Xa, whereas *Kdm5c* escaped XCI and showed diallelic expression (Fig. 3.1B, mRNA tracks). These expression patterns correlated with a differentially methylated region (DMR) at the promoter of *Kantr* that is hypermethylated at CG sites (repressed) on Xi (Fig. 3.1B, box 1). In contrast, we observed CG hypomethylation on both Xi and Xa in the CpG island at the promoter of the escape gene *Kdm5c*, as expected [29] (Fig. 3.1B, box 2).

Gene body mCH has been associated with transcriptional repression in mammalian brain cell types [9, 4, 5]. Consistent with this repressive association, mCH on Xa is highest in the intergenic region upstream of *Iqsec2* and relatively lower in gene bodies of expressed genes *Iqsec2*, *Kdm5c*, and *Kantr*. This pattern is similar to the distribution of mCH on the male X ($r = 0.94$). In contrast, mCH on Xi presents an opposite (positive) correlation with transcription: Xi is remarkably void of mCH except in the gene body of the escape gene, *Kdm5c*, where mCH

is enriched (3.09%) and more abundant than on Xa (0.26%; $p = 0.001$).

Allele-specific mCH is also evident within the X-inactivation center (XIC), a 10- to 20-Mb region that controls the establishment and maintenance of XCI [98, 104]. As expected, *Xist*, the long noncoding RNA (lncRNA) that triggers inactivation in cis, is monoallelically expressed from Xi; the escape gene, *Ftx*, is diallelically expressed; and *Chic1* is monoallelically expressed from Xa (Fig. 3.1C). Promoter mCG is consistent with this pattern of expression: *Chic1* is hypomethylated on Xa, *Xist* is hypomethylated on Xi, and *Ftx* is hypomethylated on both Xa and Xi (no DMR) (Fig. 3.1C, boxes 1-3). Gene body mCH is relatively high throughout the XIC on Xa, particularly in bodies of unexpressed genes (*Cdx4*, *Tsx*, *Tsix*, and *Xist*), and lowest in bodies of expressed genes (*Chic1* and *Ftx*). Once again, this pattern is very similar to male X ($r = 0.96$). By contrast, mCH on Xi is associated with transcriptional activity. It is enriched throughout a region upstream of *Xist* that includes the escape genes *Jpx* and *Ftx* and, to a lesser extent, within *Xist* itself.

At the XIC, allele-specific regulation of expression on the X chromosome is maintained, in part, through physical segregation of epigenetically defined chromatin regions called topologically associated domains (TADs) [105]. We found that the start of the mCH-enriched region upstream of *Xist* aligns precisely with the boundary between two TADs identified by $\sim$ 20-kb-resolution chromosome conformation capture [105]. We further examined this correspondence throughout the $\sim$ 5-Mb region surrounding *Xist* and found an additional block of enriched mCH on Xi (Appendix B.1). This region coincides with the boundaries of a TAD comprising two escape genes, *Pbdc1* and *5530601H04Rik* [105] (Appendix B.1). This correspondence suggests that mCH accumulates within topologically defined domains of accessible chromatin (i.e., active TADs) on Xi.

These results demonstrate mCH at escape genes, and the XIC is positively correlated with expression from Xi, in contrast to the repressive association observed for both mCG and mCH on Xa and autosomes.

3.2.3 Differential mCH and mCG Between Xa and Xi Predict Escape Genes

To examine the relationship between escape genes and DNA methylation more broadly, we profiled our MethylC-Sequencing (Seq) and RNA-Seq data across all X-linked genes. The distribution of RNA, mCG, and mCH was dramatically different on Xi and Xa (Appendix B.2). First, we identified escape genes in mouse frontal cortex using a binomial model that detects genes with a significant proportion of mRNA-Seq reads from Xi (Methods). In all, we found 11 genes that escaped inactivation (Fig. 3.2A and Table 1). Nine of these escape genes (*Xist*, *Ddx3x*, *Kdm6a*, *Kdm5c*, *Eif2s3x*, *5530601H04Rik*, *Ftx*, *Slc16a2*, and *Gpm6b*) are consistent with a previous survey in whole mouse brain samples [99]. In addition, we detected diallelic expression of *Tceal5* and *Gpr34*, suggesting they may be novel escape genes in frontal cortex.

We then compared gene body mCH between Xa and Xi for all X-linked genes. On Xa, the median mCH level of gene bodies was 0.88% (range: 0.00-4.53%), and the pattern was similar to the male X ($r = 0.93$; Appendix B.3). In contrast, mCH was statistically undetectable on Xi within the majority of gene bodies covered by our data (980 genes; Fig. 3.2B). In all, we identified 13 genes with statistically significant gene body mCH on Xi (Table 1). These genes included seven known escape genes (*Ddx3x*, *Xist*, *Eif2s3x*, *Kdm5c*, *Kdm6a*, *5530601H04Rik*, and *Ftx*), representing a statistically significant overlap ($p < 10^{-19}$, hypergeometric test). In addition, three genes with gene body mCH on Xi had significant expression from Xi in one of our two replicates and have been previously reported as escape genes in whole brain (*Firre* and *Pbdc1*) [99] or eye (*Tmem29*) [106]. One other gene, *Jpx*, with mCH enrichment is located within the XIC and plays a direct role in XCI but is not significantly expressed from Xi. Finally, we identified significant mCH at the lncRNA *4933407K13Rik*, which is expressed from the macrosatellite locus *Dxz4*, a region that binds CTCF on Xi and plays a role in organizing the topology of Xi [107, 108].

In contrast to the enrichment of mCH on Xi at gene bodies of escape genes, we found strong depletion of mCG at the promoters of many of these genes (Fig. 3.2C) as previously observed for human escape genes in blood [29]. Whereas the promoters of most X-linked

genes are marked by increased mCG on Xi, seven of the escape genes showed significant hypomethylation on Xi. *Pbdc1* was also significantly hypomethylated on Xi, consistent with being an escape gene as previously reported [99]. *Xist* was the only escape gene hypermethylated at its promoter on Xa, which is consistent with its silencing on Xa.

Integrating our findings for mCG, mCH, and gene expression (Fig. 3.2D), we observe the following pattern: Seven escape genes (*Ddx3x*, *Xist*, *Eif2s3x*, *Kdm5c*, *Kdm6a*, *5530601H04Rik*, and *Ftx*), and possibly *Pbdc1*, have distinct methylation patterns with hypomethylated CG promoters and hypermethylated CH gene bodies, and four other escape genes (*Slc16a2*, *Gpm6b*, *Gpr34*, and *Tceal5*) have Xi methylation patterns similar to nonescape genes with CG hypermethylation and CH hypomethylation. Considering prior surveys of escape from XCI [99], we observe that all CG-hypomethylated and CH-hypermethylated genes escape XCI across multiple tissues. *Gpm6b* was reported to escape XCI only in brain and lacks the unique DNA methylation signatures we observed at genes that ubiquitously escape XCI. Comparisons of DNA methylation and expression levels show that escape genes form a highly distinctive compartment in which a relative increase in mCH on Xi compared with Xa marks genes that escape X inactivation (Fig. 3.2D).

3.2.4 Analyses of Intergenic Regions

To include intergenic regions in our analysis, we next examined DNA methylation in 2.5-kb bins across the X chromosome. Fig. 3.3A shows the location of escape genes on the X chromosome (triangles), of significantly methylated bins on Xi (blue ticks), and of genes identified as CH-hypermethylated (stars), highlighting the chromosome-wide distribution of these genes. On Xi, mCH is absent throughout most intergenic regions, punctuated by 12 significant peaks of enriched mCH corresponding to the previously identified CH-methylated genes (Fig. 3.3B). In contrast, mCG is high throughout Xi, with a few exceptions corresponding to escape genes and the XIC (Fig. 3.3D). To quantify these patterns, we used MethylSeekR [109] to call CG unmethylated regions (CG-UMRs), which typically correspond to promoters of expressed

genes. We identified 437 significant CG-UMRs on Xa but only 37 on Xi [false discovery rate (FDR) < 0.05, coverage by five or more reads in $\leq 30\%$ of CG sites; Dataset S1]. Thirty-two CG-UMRs on Xi correspond to genes located in the XIC and to escape genes.

We noted that 14% of CG-UMRs on Xi fall within or proximal to *Bcor*, which does not appear to escape in brain based on our RNA-Seq analyses but was listed as an escape gene in a cell line in a previous study [99] (Fig. 3.3D, arrowhead and Appendix B.4). *Bcor*, a gene in which mutations can lead to oculofaciocardiodental (OFCD) syndrome, was previously shown to be half as methylated in females (Xa + Xi) as in males in human blood and buccal tissue [110]. This pattern runs counter to the pattern at other nonescape genes, where males show lower mCG compared with females, suggesting a unique pattern at *Bcor*. Our results reveal that *Bcor* contains multiple CG clusters hypomethylated on Xi on the paternal allele. This finding is consistent with the previous finding in humans, suggesting a conserved epigenomic pattern. Wamstad et al. [111] suggested that *Bcor* is unlikely to be a maternally expressed imprinted gene because the observed mother-to-daughter transmission of *Bcor* mutations in OFCD is not lethal. Here, we further reason that *Bcor* is unlikely to be imprinted to express only the paternal allele because we should not observe a phenotype in mother-to-daughter transmission if the maternal allele is not expressed. Therefore, the allele-specific methylation observed at *Bcor* is most likely specific to the activation state of the chromosome rather than the parental origin.

3.2.5 Allele-Specific Methylation and Imprinting

In addition to its role in XCI, allele-specific DNA methylation plays a key role in regulating autosomal imprinted regions. A previous study profiled DNA methylation using MethylC-Seq in brain samples from male Cast/129 F1 hybrid mice and identified imprinted autosomal methylation in both CG and CH contexts [9]. The C57/ *spretus* F1 female mice in our study have twice as many SNPs (41.7 million compared with around 20 million for Cast/129 hybrids). Although our *Xist* mutant mouse line could not be used to produce a reciprocal cross (i.e., maternal *spretus*

x paternal C57) to distinguish species-of-origin vs. parent-of-origin effects, we nevertheless analyzed maternal vs. paternal differences in methylation at known imprinted regions to confirm and extend prior observations using a different genetic background. We compared maternal and paternal mCG levels at promoters of all autosomal genes (Fig. 3.4A). As expected, most genes have equal mCG levels on the two alleles. We then identified genes with allelic differences in mCG (allele1 $> 75\%$ and allele2 $< 25\%$) and a significant DMR in the promoter. Our results recapitulate the imprinted loci previously identified [9]. In addition, we found maternal mCG at the imprinted *Nnat* promoter, a gene Xie et al. [9] could not examine due to a lack of SNPs in their cross.

We next sought to connect allelic differences in methylation with expression. We found 77 autosomal genes that were differentially expressed between alleles in both replicates (FDR < 0.05 and log2-fold change > 1.5). Differential expression of these genes was significantly correlated with allelic differences in both promoter mCG ($r = -0.558$, $p < 1e-6$) and gene body mCH ($r = -0.312$, $p = 0.0027$) (Fig. 3.4B). Focusing on DMRs previously reported to be imprinted in a parent-of-origin-dependent manner using reciprocal crosses [9], we found largely consistent mCG differences. Several imprinted DMRs identified in the Cast/129 F1 hybrids (*Casc1* intragenic, *6330408a02Rik* 3' end, *FR149454* promoter, *FR085584* promoter, *Myo10* intragenic, *Vwde* promoter, and *Pvt1* promoter) fail to show allele-specific CG methylation in our data, suggesting they might not be conserved across mouse species (Fig. 3.4C and Dataset S2).

Whereas CG DMRs were localized to discrete regions (900-bp median size), we found substantial allele-specific differences in autosomal mCH that extended over much larger domains encompassing one or more gene bodies, as observed previously [9]. Surprisingly, we found that allele-specific mCH could exhibit either the same asymmetry as allele-specific mCG or the reverse asymmetry. For example, the imprinting control region for the *Kcnq1* gene, located at the promoter of the antisense transcript *Kcnq1ot1*, is a 2.5-kb DMR with allele-specific mCG on the maternal allele. However, there is a much larger mCH DMR, spanning the entire *Kcnq1ot1*

transcript (87.6 kb), that is also hypermethylated on the maternal allele (Fig. 4D). A reciprocal example is the maternally imprinted locus on chromosome 12 containing *Meg3*, *Rian*, and *Mirg*, where the paternal allele is marked by discrete allele-specific mCG and more diffuse mCH (Fig. 3.4E).

In contrast, we observed the reverse asymmetry (i.e., lower mCG on the paternal allele and lower mCH on the maternal allele) at the $\sim$ 3.6-MB region of chromosome 7 containing imprinted *Snrpn*, *Snurf*, and *Magel2* (Fig. 3.4F). Genetic variants in this region can cause Prader-Willi or Angelman syndrome, depending on which allele is affected. This locus contains a large CH DMR spanning $\sim$ 3.6 MB that is hypomethylated on the maternal allele, whereas CG hypomethylation is restricted to the paternal allele and occurs mainly at the promoters of imprinted genes within the locus. Another example of a reverse asymmetry between allele-specific mCG and mCH occurs within the *Nesp*/*Gnas*/*Nespas* locus (Appendix B.5). Together, these patterns of allele-specific autosomal mCG and mCH suggest a complex relationship between the two types of methylation, with both positive and negative correlations.

As with X inactivation, our analysis of methylation at imprinted autosomal loci reveals that mCG and mCH have contrasting allele-specific distributions indicating at least partly independent roles in gene regulation.

3.3 Discussion

Allele-specific regulation of domains of active and inactive chromatin is critical for healthy brain development in mammals, yet the landscape of DNA methylation within these domains has largely been studied without allele-specific resolution. Using MethylC-Seq and RNA-Seq in the frontal cortex of female transgenic mice with deterministic XCI, we obtained allele-specific, base-resolution DNA methylation and transcription profiles. In all, we identified 11 genes escaping XCI. Methylation profiling showed that the Xi chromosome was largely devoid

of CH methylation, whereas most gene promoters showed CG hypermethylation. Seven escape genes (plus *Pbdc1*, a previously reported escape gene that escaped in one of our replicates) showed a pattern of CH hypermethylation in gene bodies and CG hypomethylation at promoters on Xi. Findings of hypo-mCG at promoters of escape genes are consistent with previous studies that analyzed CG methylation across multiple tissues in humans [29, 112]. However, mCH had not been previously examined on Xi and Xa. Although the only genes with hypo-mCG on Xi were also CH-hypermethylated, there were three additional genes with only hyper-mCH, suggesting distinct roles for the DNA methylation in these contexts.

mCH accumulates during postnatal development of frontal cortical neurons, reaching high levels in the adult mouse and human brain [9, 4]. Indeed, the abundance of mCH is comparable to the abundance of mCG in adult neurons, and mCH is found in both excitatory and inhibitory neuron types [5]. Our findings show that mCH is a high-fidelity epigenomic marker of allele-specific active chromatin domains, such as genes escaping X inactivation, which can be used for functional genomic annotation. However, the functions of mCH, if any, are unknown [113]. Promoter mCG and gene body mCH are associated with transcriptional repression and are generally correlated. Our findings in genomic regions affected by XCI and parental imprinting demonstrate a partial dissociation between the CG and CH contexts of DNA methylation.

First, we identified hypo-mCG and hyper-mCH at a subset of escape genes. Most of these escape genes have been reported to escape across multiple tissues, so it is unclear if the presence of mCH, which is specific to the brain, is necessary for escaping inactivation or if it is a consequence of chromatin accessibility. Second, we identified three genes with hyper-mCH on Xi that lacked allele-specific CG hypomethylation. Because these cells are postmitotic and there is no known active mechanism for removing mCH, the presence of mCH at these regions may serve as a marker of previously active chromatin. Tissue type-specific or cell type-specific differential mCG in adult cells has been shown to reflect early developmental processes at so-called vestigial DMRs [5, 114]. Third, we examined numerous imprinted genes and identified a complex and variable

relationship between mCG and mCH. Although CG and CH sites are often hypomethylated on the same allele, as in the case of *Kcqn1ot1* and *Meg3*, we also identified a striking pattern in the Prader-Willi/Angelmann syndrome-associated region of chromosome 7, where allele-specific mCG and mCH were oppositely regulated.

A previous study that identified escape genes in whole brain reported 17 escape genes [99], and nine of the escape genes we found in frontal cortex overlap with these results. Our results demonstrate the presence of a DNA methylation signature at the large majority of escape genes. The absence of a methylation signature at other escape genes suggests there may be more than one pathway to escape from XCI. For example, *Gpm6b* has been previously reported to escape inactivation only in the brain. The characteristic escape gene hypo-mCG, which presumably occurs early in development in a precursor cell type, would be inconsistent with its escape only in brain. Therefore, there may exist a brain-specific mechanism to support the later escape of *Gpm6b*. In addition to tissue specificity, our results may suggest temporal dynamics to escape genes in the brain, as has been previously reported in mouse embryo [115]. If mCH on Xi indeed marks active chromatin, then the nonescape genes with mCH may escape earlier in development and be downregulated in adults. Alternatively, it is possible these genes are expressed from Xi continuously over brain development at levels too low to achieve significance in our analysis.

Finally, our findings also contribute to the evidence of sex differences in the epigenome [116, 117]. Sex chromosomes and XCI are genetic drivers of sex differences that, together with the effect of sex hormones, result in sexually dimorphic brain structure and cognition. Sex-specific DNA methylation can help us to understand how genes are regulated differentially between sexes and to develop efficacious treatments for disease in both sexes, an important objective set forth by the NIH [118]. To this point, our analysis shows that DNA methylation on female Xa and male X is largely similar in both CG and CH contexts (Appendix B.3). The differences we identified here are specific to Xi and support a role for escape genes in the development of sex differences, particularly in humans, where escape genes are far more numerous than in mice [119]. The

results may also shed light on the nonverbal learning disabilities and attention deficit hyperactivity disorder-like symptoms exhibited in Turner syndrome (45,X) females missing the Xi [28]. If differences between typical females and those females with Turner syndrome are restricted to escape genes and the nonescape genes with hyper-mCH, then these genes likely play a critical role in healthy female brain development.

3.4 Methods

3.4.1 Mouse Model/Animals

Xist is an lncRNA that initiates inactivation *in cis*. Previous work has shown that deletion of a proximal A repeat inhibits *Xist* transcription and prevents inactivation [30]. To profile allele-specific DNA methylation and transcription in a deterministic model of XCI, we used 14-wk-old female F1 progeny of C57BL/6 Xist mutant female mice and *M. spretus* wild-type male mice (The Jackson Laboratories) [99]. Due to the deletion, the maternal X chromosome (C57) failed to inactivate and was the Xa in all cells, whereas the paternal chromosome (*spretus*) was ubiquitously inactivated (Xi). Furthermore, the genetic variability between the two mouse species allowed us to assign sequencing reads to the parent of origin. We collected samples from four biological replicates at 14 wk of age. Animals were weaned in groups of three to five per cage. Female F1 pups were genotyped at weaning to confirm the presence of the mutant allele.

Animals were anesthetized with CO2, followed by cervical dislocation. Brains were removed and rinsed in cold PBS. For dissection, whole brains were placed in cold DMEM supplemented with 10% (vol/vol) FBS. The prefrontal cortex (PFC) was obtained by first removing the cerebellum, followed by slicing coronally 1 mm at the bregma and carefully isolating the frontal cortical tissue under a dissecting microscope. PFC samples were rapidly frozen on dry ice until processing. DNA and RNA were isolated from pooled PFC samples obtained from two individuals from separate litters. All protocols were approved by the University of Washington's

Institutional Animal Care and Use Committee.

3.4.2 MethylC-Seq

Libraries were sequenced as single-end reads and prepared using the following procedure: Genomic DNA was extracted from ground, frozen tissue using the DNeasy Blood and Tissue Kit (Qiagen, Valencia, CA). Two micrograms of genomic DNA was spiked with 10 ng of unmethylated cl857Sam7 Lambda DNA (Promega). The DNA was fragmented with a Covaris S2 instrument to 150-200 bp, followed by end repair and addition of a 3' A base. Cytosine-methylated adapters provided by Illumina were religated to the sonicated DNA at 16 °C for 16 h with T4 DNA ligase (New England Biolabs). Adapter-ligated DNA was isolated by two rounds of purification with AMPure X P beads (Beckman Coulter Genomics). Adapter-ligated DNA ($\leq$450 ng) was subjected to sodium bisulfite conversion using the Methyl Code Kit (Life Technologies) as per the manufacturer's instructions. The bisulfite-converted, adapter-ligated DNA molecules were enriched by four cycles of PCR with the following reaction composition: 25 μL of Kapa Hi Fi Hotstart Uracil+Readymix (Kapa Biosystems) and 5 μL of TruSeq PCR Primer Mix (Illumina) (50 μL final). The thermocycling parameters were 95 °C for 2 min; 98 °C for 30 s; and then four cycles of 98 °C for 15 s, 60 °C for 30 s, and 72 °C for 4 min, ending with one 72 °C 10-min step. The reaction products were purified using AMPure X P beads. Two separate PCR reactions were performed on subsets of the adapter-ligated, bisulfite-converted DNA, yielding two independent libraries from the same biological sample for subsequent sequencing using a HiSeq 2500 system (Illumina).

3.4.3 mRNA-Seq Library Preparation

Ribosomal RNA was removed from samples using a Ribo-Zero rRNA Removal Kit (Illumina). mRNA-Seq libraries were then generated using the TruSeq Stranded RNA LT Kit

(Illumina) according to the manufacturer's instructions. Samples were sequenced using the HiSeq 2500 system.

3.4.4 Reference Genomes

The mm10 reference genome is the reference for the C57BL/6J strain. For *M. spretus*, we created a pseudo-reference genome by updating the mm10 reference with known C57-*spretus* SNPs as reported by the Sanger Institute [120] (www.sanger.ac.uk/science/data/mouse-genomes-project). We only retained high-confidence SNPs that passed all quality filters (denoted in the file as FI = 1), resulting in $\sim$ 1.95 million SNPs on chromosome X. Before allele sorting, our reads covered 89.1% of the genome. We were able to assign 68.6% of reads to one of the alleles, yielding broad and deep coverage for C57 (78.2% covered, 11.93 average read depth) and *spretus* alleles (70.1% covered, 9.79 average read depth). High coverage (at least five reads) was achieved at 67.0% of the genome in C57 and at 51.6% of the genome in *spretus*.

3.4.5 Mapping of MethylC-Seq Data

Sequencing reads were mapped separately to both the C57 and *spretus* reference genomes using Methylpy [4]. Unmethylated phage lambda DNA was spiked into each sequencing run, allowing us to estimate the bisulfite nonconversion rate directly (0.36% and 0.40% for the two replicates, respectively). Reads that mapped to one or both of the reference genomes were then pooled and assigned to the parent of origin, corresponding to Xa (C57) and Xi (*spretus*). Only reads containing one or more SNPs that matched 100% to one parental reference or the other were retained and used in the analysis. We noted that a small proportion ($\sim$2.5%) of CH sites were covered by reads that contained sequence mismatches potentially consistent with a CG position. To prevent contamination from these ambiguous sites, we excluded them from our analysis.

3.4.6 Mapping of mRNA-Seq Data

The mRNA-Seq data were mapped as previously reported [99]. The mRNA-Seq reads were mapped separately to both the C57 and *spretus* reference genomes using TopHat2 [121] with default parameters. The transcriptome included only exons as defined in the GENCODE release M7 (level 3) comprehensive gene annotation file (www.gencodegenes.org). High-quality reads (mapping quality score MAPQ ≥ 30) that mapped to one or both of the reference genomes were then pooled and assigned to the parent of origin, corresponding to Xa (C57) and Xi (*spretus*). Only reads containing SNPs that matched 100% to one parental reference or the other were retained for analysis.

3.4.7 Data Analysis

Browser representations were created using Anno-J (www. annoj.org) [10]. Pearson correlations were used except where stated otherwise. P values were < 0.01 unless otherwise stated.

3.4.8 MethylC-Seq Analysis

Methylation was analyzed separately for the CG and CH (i.e., CA, CC, CT) contexts. We examined mCG at promoters and mCH in gene bodies, both of which correlate with transcriptional repression in the brain [9, 4, 5]. Gene transcription start and end sites were taken from GENCODE release M7 (level 3), and promoters were defined as $\pm 1,000$ bases from the transcription start site. Methylation was quantified as the number of methylated cytosine base calls (m) divided by total cytosine base calls (c), and was corrected for the nonconversion rate (NCR; calibrated using spike-in lambda DNA) using the maximum likelihood formula:

$$mC = g\left[\frac{m/c - NCR}{1 - NCR}\right]$$

$$g[x] = max[x, 0]$$

Importantly, MethylC-Seq cannot distinguish between methylcytosine and hydroxymethylcytosine, which is present at significant levels in brain in the CG context [4, 56]. However, prior studies using Tet-assisted bisulfite sequencing have shown that there is no detectable hydroxymethylation at non-CG sites [4, 60]. Therefore, although differences in CG methylation could be driven by changes in one or more types of methylation, our analyses regarding CH methylation are not affected by this ambiguity.

3.4.9 mRNA-Seq Analysis

Identification of escape genes using mRNA-Seq adhered to previously published methods [99]. First, reads that mapped to C57, *spretus*, or both reference genomes were aggregated and used to quantify diploid expression in fragments per kilobase of transcript per million mapped reads (FPKM) with Cufflinks [122]. Next, reads were assigned to the Xa or Xi only if all SNPs within a read corresponded to either the C57 (Xa) or *spretus* (Xi) reference. Reads that did not meet these criteria or that contained no SNPs were discarded. We then quantified haploid expression as allele-specific reads per 10 million mapped reads (SRPM). Finally, a binomial model was used to compute a confidence interval for the expression of each gene on Xi [99]. A gene was said to escape inactivation significantly if diploid expression FPKMs were $\geq$1, Xi-SRPM was $\geq$1, and the lower bound of the 99% confidence interval from the binomial model was > 0.

3.4.10 Definition of CH-Hypermethylated and CG-Hypomethylated Genes

We modeled the methylation of genes on Xi using a mixture distribution that we fit using an iterative procedure. We first fit a beta-binomial distribution for the apparent CH methylation levels of all gene bodies on the inactive Xi by maximum likelihood. We then used this beta-binomial

distribution to compute a p value for each gene (i.e., the likelihood of observing that gene's mCH level, given the null distribution) and marked any genes with significantly greater mCH (FDR < 0.05 using the Benjamini-Yekutieli correction) as "hyper-mCH" genes. We then repeated our fitting procedure using only genes that were not marked as hyper-mCH. This procedure was repeated until it converged on a set of hyper-mCH genes. Only genes with significant mCH in both replicates were reported in our results. The same analysis was applied on CG methylation in promoters to identify significantly CG hypomethylated genes on Xi.

3.4.11 Additional Datasets

MethylC-Seq data for 6-wk-old male mouse frontal cortex and fetal brain tissue (embryonic day 13.5, mixed male and female) were from a previously published study [4]. These datasets were mapped to mm10 and processed using the same methods described above.

3.4.12 Data Access and Browser

Data are accessible in the Gene Expression Omnibus (GEO) database (accession no. GSE83993). Data are also displayed via a web-based browser at brainome.ucsd.edu/mm_xist_hybrid.

3.5 Acknowledgements

Chapter 3, in full, is a reprint of the material as it appears in *Proceedings of the National Academy of Sciences* 2017. Christopher L. Keown, Joel B. Berletch, Rosa Castanon, Joseph R. Nery, Christine M. Disteche, Joseph R. Ecker, and Eran A. Mukamel. The dissertation author was the primary investigator and first author of this paper.

Table 3.1: Escape genes and CH-hypermethylated genes

Gene	Gene body, % mCH		Gene promoter, % mCG		Diallelic RNA abundance, FPKM	Allele-specific RNA abundance, SRPM	
	Xa	Xi	Xa	Xi		Xa	Xi
Escape genes with mCH on Xi							
5530601H04Rik	0.19	***2.40***	3.7	***2.1***	6.49	36.9	***12.3***
Ddx3x	0.17	***1.89***	0.2	***6.3***	44.8	84.9	***44.5***
Eif2s3x	0.10	***1.67***	3.4	***1.4***	22.5	30.5	***24.4***
Ftx	0.42	***2.47***	11.6	***10.1***	6.05	67.0	***7.65***
Kdm5c	0.28	***3.11***	1.2	***3.0***	7.40	57.9	***25.0***
Kdm6a	0.39	***2.22***	0.4	***0.02***	5.43	22.7	***8.25***
Xist	1.79	***0.79****	86.0	***0.5***	24.9	5.74	***546***
Firre	0.94	***1.16***	7.8	73.6	6.25	64.2	***1.91****
Pbdc1	0.52	***3.27***	3.7	***2.1***	4.40	17.2	***4.02****
Tmem29	0.43	***1.52***	6.1	27.8	10.2	28.8	***3.52****
Escape genes without mCH on Xi							
Gpm6b	0.57	0.04	13.7	79.1	121	252	***27.2***
Gpr34	1.15	0.04	65.9	59.8	3.62	13.9	***4.01***
Slc16a2	1.15	0.16	0.1	74.3	5.31	34.1	***3.91***
Tceal5	0.00	0.12	33.2	67.9	16.7	23.5	***9.93***
Other CH-hypermethylated genes							
Gm38020[†]	0.07	***3.85***	93.5	93.4	2.76	43.7	4.99
Jpx[†]	0.77	***3.46***	6.8	26.5	1.21	7.51	0.815
4933407K13Rik[‡]	0.25	***1.05***	5.9	51.2	0.565	8.40	0.10

Bold italic values indicate significant gene body mCH on Xi, significant CG promoter hypomethylation on Xi, or Xi RNA abundance (FDR < 0.05 in both replicates).

*Genes significant in one replicate only.

[†]Occurs within the XIC. *Gm38020* overlaps the escape gene *Ftx*, and *Jpx* has previously been reported to escape (42).

[‡]Located at the *Dxz4* macrosatellite, which is involved in Xi chromosome topology.

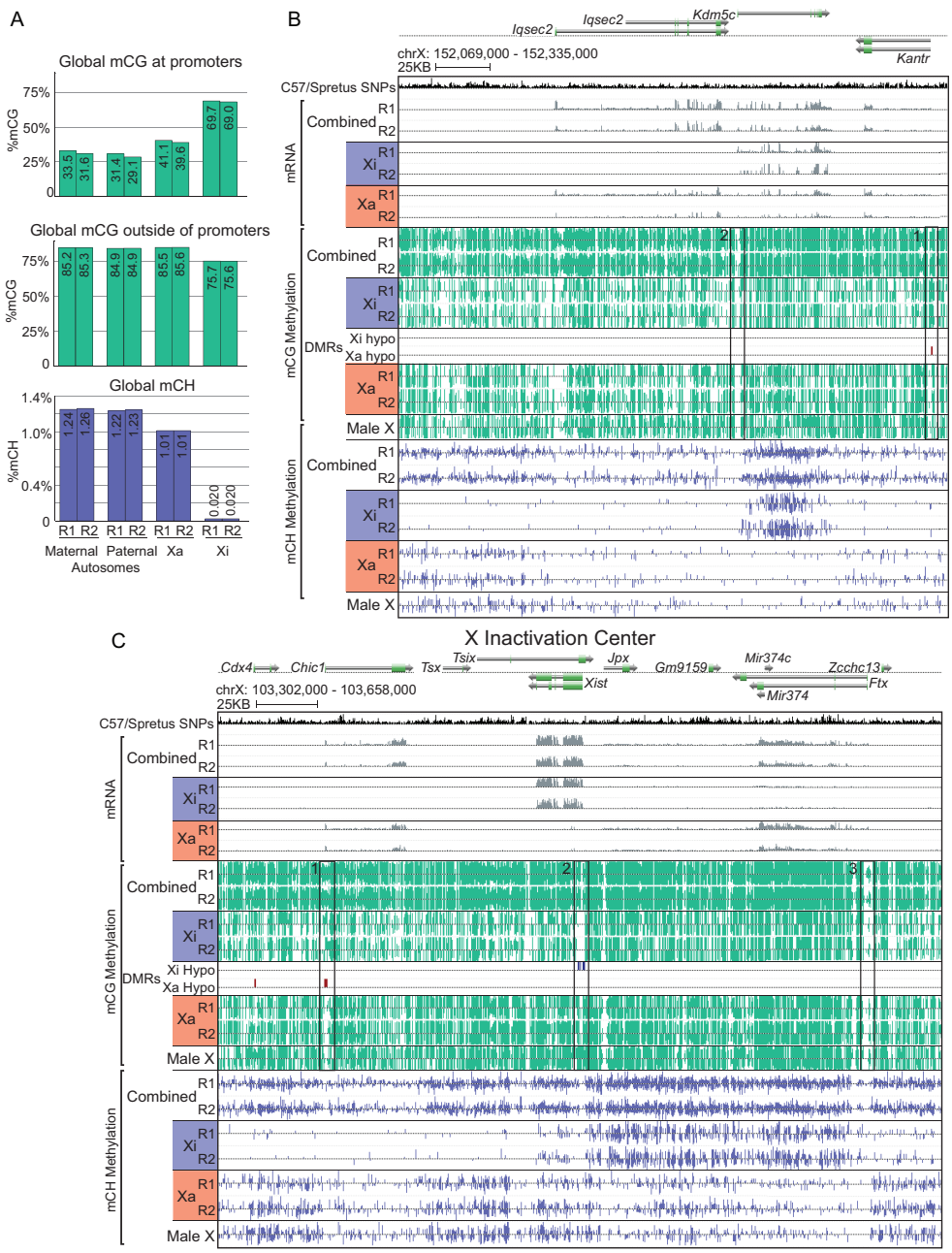

Figure 3.1: Ultrasparse mCH on Xi correlates with escape domains. (**A**) Allele-specific mCG and mCH levels on autosomes and chromosome X. Browser view of methylation and expression for the *Kdm5c* locus (**B**) and the XIC (**C**). Ticks show the methylation level at individual cytosine positions (CG, green; CH, blue) on the forward (upward ticks) and reverse (downward ticks) strands. Combined tracks show both alleles, whereas the Xa and Xi tracks include only reads sorted using SNPs between C57 and *spretus*. Monoallelically expressed genes (*Iqsec2* and *Chic1*) and intergenic regions harbor mCH on Xa only, whereas diallelically expressed escape genes (*Kdm5c*) and the Xi-expressed noncoding RNA Xist contain dense mCH on Xi. Male X data are from 6-wk-old frontal cortex **??**. chrX, chromosome X; R1, replicate 1; R2, replicate 2.

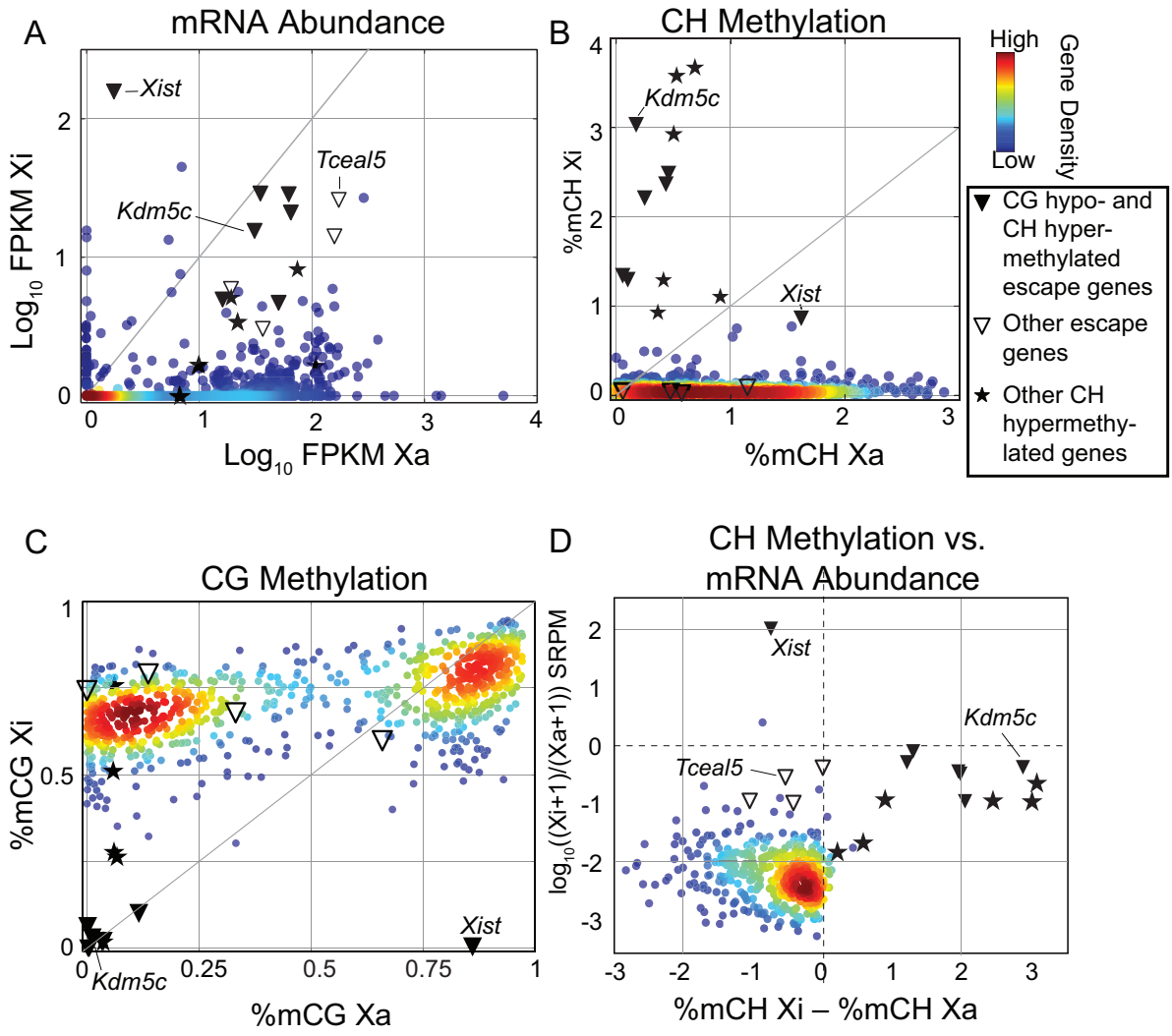

Figure 3.2: Escape genes are marked by unique mCG and mCH patterns. (**A**) Allele-specific expression for each gene on chromosome X demonstrates nonrandom X inactivation and identifies escape genes. Each point corresponds to a gene, and the color of the point indicates the density of genes. Genes with a significant number of reads originating from Xi in both replicates are indicated as escape genes (Table 1). (**B**) mCH in gene bodies is virtually absent from Xi, except at a subset of escape genes, genes involved in X inactivation (*Jpx*), and a few additional loci. (**C**) Most gene promoters harbor dense mCG on Xi, whereas the bulk of escape genes have unmethylated promoters. A subset of escape genes (*Tceal5*, *Gpr34*, *Slc16a2*, and *Gpm6b*) violate this pattern and contain high levels of promoter mCG. (**D**) Comparison of gene body mCH (x axis) and expression (y axis) between Xa and Xi reveals the clustering of escape genes.

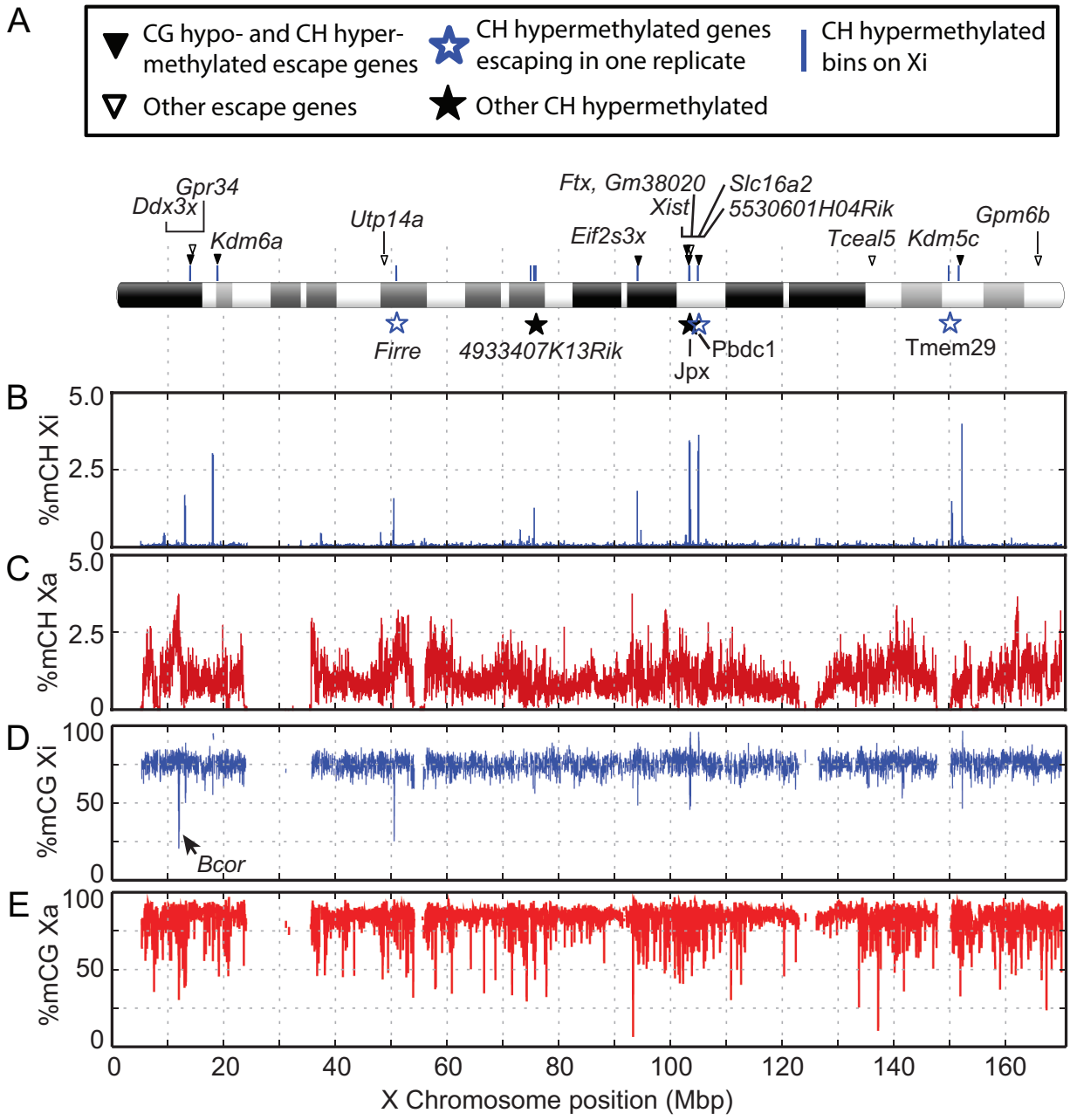

Figure 3.3: X-chromosome-wide landscape of allele-specific DNA methylation. (**A**) Locations of genes escaping X inactivation, mCH-enriched regions (blue ticks), and mCH-enriched genes. (**B** and **C**) mCH is abundant across Xa but sparsely distributed on Xi. (**D** and **E**) mCG is abundant across both Xi and Xa but locally depleted at promoter regions of nonescape genes on Xa.

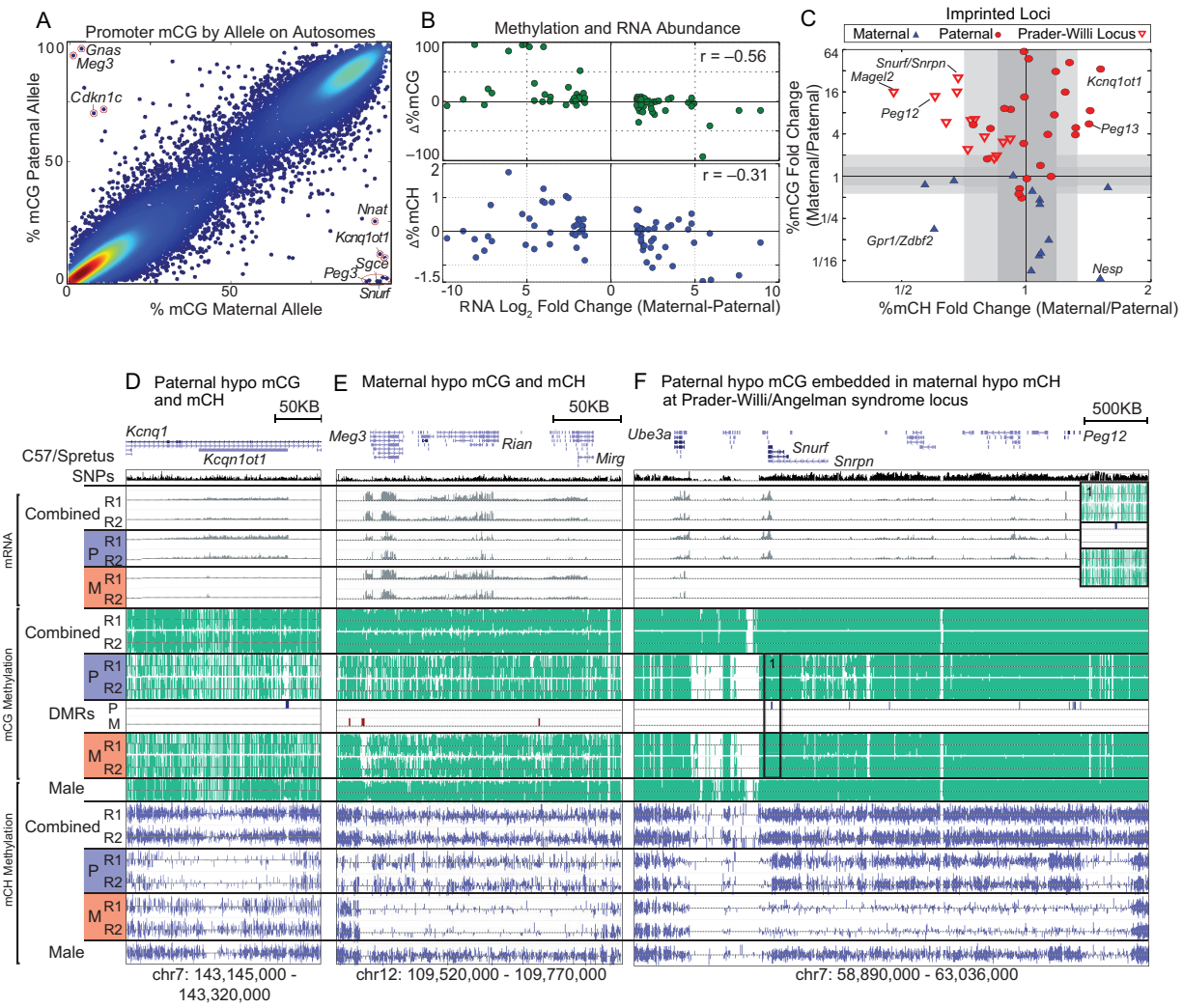

Figure 3.4: Imprinted genes marked by allele-specific mCG and mCH. (**A**) Allelic differences in promoter mCG in autosomal genes recapitulate previously identified imprinted genes [9], indicated by red circles, including maternally imprinted *Peg13* and paternally imprinted *Gnas*, *Meg3*, and *Cdkn1c*, and suggest that *Nnat* has allele-specific methylation or is imprinted. (**B**) Eighty autosomal genes with allelic differential expression ($> 1.5 log_2$-fold change) are plotted against their allelic differences in mCG and mCH. (**C**) Differential CG methylation for imprinted DMRs identified by Xie et al. [9] is compared with differential mCH in the surrounding region ($\pm$10 kb). Shaded regions indicate 95% (dark shade) and 99% (light shade) confidence intervals for a null distribution obtained by comparing allelic differences in methylation across chromosomes in all autosomes in mCH using 20-kb bins (x axis) and in mCG using 1-kb bins (y axis). Browser views of imprinted loci *Kcnq1ot1* (**D**); *Meg3*, *Rian*, and *Mirg* (**E**); and *Snrpn*, *Snurf*, and *Magel2* (**F**). (Inset) Zoomed-in view of paternal hypo-mCG DMR at the promoter of *Snurf*. M, maternal; P, paternal.

Chapter 4

Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex

4.1 Article

Mammalian neuron types are identified by their structure, electrophysiology, and connectivity [123]. The difficulty of scaling traditional cellular and molecular assays to whole neuronal populations has prevented comprehensive analysis of brain cell types. Sequencing mRNA transcripts from single cells or nuclei has identified cell types with unique transcriptional profiles in the mouse brain [36, 37] and human brain [39]. However, these methods are restricted to RNA signatures, which are influenced by the environment. Epigenomic marks, such as DNA methylation (mC), are cell type-specific and developmentally regulated, yet stable across individuals and over the life span [5, 124, 4]. We theorized that epigenomic profiles using single-cell DNA methylomes could enable the identification of neuron subtypes in the mammalian brain.

During postnatal synaptogenesis, neurons accumulate substantial DNA methylation at

non-CG sites (mCH) and reconfigure patterns of CG methylation (mCG) [4]. Patterns of mCG and mCH at gene bodies, promoters, and enhancers are specific to neuronal types [5, 124, 4, 125]. Gene body mCH is more predictive of gene expression than mCG or chromatin accessibility [82]. Because mCH is modulated over large domains, single-neuron methylomes with sparse coverage can be used to accurately estimate mCH levels for more than 90% of the genome by using coarse-grained bins (100 kb) (Appendix C.1). Whereas single-cell RNA sequencing mainly yields information about highly expressed transcripts, single-neuron methylome sequencing assays any gene or nongene region long enough to have sufficient coverage.

We developed a protocol for single-nucleus methylcytosine sequencing (snmC-seq) and applied it to neurons from young adult mouse (age 8 weeks) and human (age 25 years) frontal cortex (FC) (Fig. 4.1 1A). snmC-seq provides a high rate of read mapping relative to published protocols [126, 127, 81] and allows multiplex reactions for large-scale cell type classification (Appendix C.2). Like other bisulfite sequencing techniques [128], snmC-seq measures the sum of 5-methyl- and 5-hydroxymethylcytosines. Single neuronal nuclei labeled with antibody to NeuN were isolated by fluorescence-activated cell sorting (FACS) from human FC and from dissected superficial, middle, and deep layers of mouse FC. We generated methylomes from 3377 mouse neurons with an average of 1.4 million stringently filtered reads, covering 4.7% of the mouse genome per cell (Fig. 4.1, B and C, and table S1). We also generated methylomes from 2784 human neurons with an average of 1.8 million stringently filtered reads, covering 5.7% of the human genome per cell (Fig. 4.1, B and C, and table S2).

We calculated the mCH level for each neuron in nonoverlapping 100-kb bins across the genome, followed by dimensionality reduction and visualization using t-distributed stochastic neighbor embedding (t-SNE [129]). The two-dimensional tSNE representation was largely invariant over a wide range of experimental and analysis parameters (Appendix C.3). A substantially similar tSNE representation was obtained using CG methylation levels in 100-kb bins, which suggests that snmC-seq could be effective for cell type classification of nonbrain tissues without

high levels of mCH (Appendix C.3F).

The mammalian cortex arises from a conserved developmental program that adds excitatory neuron classes in an inside-out fashion, progressing from deep layers (L5, L6) to middle (L4) and superficial layers (L 2/3) [123]. Inhibitory interneurons arise from distinct progenitors in the ganglionic eminences and migrate transversely to their cortical locations [130]. We used mCH patterns to identify a conservative and unbiased clustering of nuclei for each species. Cluster robustness was validated by shuffling, downsampling, and comparison to density-based clustering (figs. S3 and S4) [131]. In addition, clustering was not significantly associated with experimental factors (e.g., batches; false discovery rate > 0.1, χ^2 test; Appendix C.5).

We applied identical clustering parameters to mouse and human cortical neuron mCH data and identified 16 mouse and 21 human neuron clusters (Fig. 4.2, A to D). Assuming an inverse relationship between gene body mCH (average mCH across the annotated genic region) and gene expression [4], we annotated each cluster on the basis of depletion of mCH at known cortical glutamatergic or GABAergic neuron markers (e.g., *Satb2*, *Gad1*, *Slc6a1*), cortical layer markers (e.g., *Cux2*, *Rorb*, *Deptor*, *Tle4*), or inhibitory neuron subtype markers (e.g., *Pvalb*, *Lhx6*, *Adarb2*) [123, 130, 132] (Fig. 4.2, E and F, and figs. S6 and S7). For most clusters, mCH depletion at multiple marker genes (figs. S6 and S7) allowed us to assign cluster labels indicating the putative cell type. For example, we found a cluster of mouse neurons with ultralow mCH at *Rorb* (Fig. 4.2E and Appendix C.16), a known marker of L4 and L5a excitatory pyramidal cells [132]. Combining this information with markers such as *Deptor* (Appendix C.6), which marks L5 but not L4 neurons, we labeled the cluster by species and layer (e.g., mL4 for mouse L4). Similarly, we used classical markers for inhibitory neurons such as *Pvalb* to label corresponding clusters (e.g., mPv for putative mouse Pvalb+ fast-spiking interneurons) [130]. We confirmed the accuracy of these classifications by comparison to layer-dissected cortical neurons (Appendix C.8, A and B) and coclustering with high-coverage methylC-seq data from purified populations of PV+ and VIP+ [5], as well as SST+ inhibitory neurons (Appendix C.8C). Aggregated single- neuron

methylomes showed consistent mCH and mCG profiles relative to bulk methylomes of matching cell populations (Appendix C.8, D and E, and Appendix C.15F). Neuronal cluster classification for each of the major cell subtypes in mouse and human cortex based on single-nuclei methylomes (Fig. 4.2G and Appendix C.9, A and B) was in good agreement with annotations based on single-cell RNA sequencing [36, 37, 39]. Gene body mCH was anticorrelated with expression levels for corresponding clusters (Appendix C.9, C to E), validating our mCH marker gene-based annotation.

We found a greater diversity of excitatory neurons in deep layers than in superficial layers for both mouse and human (Fig. 4.2). In both species, we identified one neuronal cluster for cortical L2/3 (mL2/3, hL2/3) and L4 (mL4, hL4), whereas L5 and L6 contained seven clusters in mouse (mL5, mL6, and mDL, where DL denotes deep-layer neurons) and 10 clusters in human (hL5, hL6, and hDL). Mouse L5 excitatory clusters (mL5-1, mL5-2) were hypomethylated at *Deptor* and *Bcl6*, which mark cortical L5a and L5b, respectively (Appendix C.6) [133]. L6 excitatory clusters included subtypes with low mCH at the L6 excitatory neuron marker *Tle4* [mL6-1, mL6-2; hL6-1, hL6-2, hL6-3 [123]]. Interestingly, several deep-layer neuron clusters (mDL-2, hDL-1, hDL-2, hDL-3) were not hypomethylated at *Tle4*. We identified marker genes for each neuron type on the basis of cell type-specific mCH depletion (table S3). Although many marker genes were either classically established [123, 132] or recently identified neuron type markers (Appendix C.10, A and B) (2-4), we identified a number of markers with no prior association to neuronal cell types (Appendix C.10, D and E, and table S3). mCH signature genes were hypomethylated in homologous clusters in mouse and human, with a few notable exceptions. For example, the mouse L5a marker *Deptor* showed no specificity for human L5 neurons (figs. S6, S7, and S10, A and B).

Most clusters were associated with classical cell type markers, but the identity of some clusters such as mDL-2 was less clear. We found that mDL-2 shares 24 marker genes with mL6-2, whereas 93 marker genes distinguish these clusters (table S3). To validate the distinction

between the two cell types, we selected a shared marker (*Sulf1*) and one unique to mL6-2 (*Tle4*) and performed double in situ RNA hybridization experiments in mouse FC (Appendix C.11). The result confirmed the mCH-based prediction of a substantial proportion of L6 neurons expressing *Sulf1* but not *Tle4*; these neurons likely correspond to mDL-2 (Appendix C.11, A to D). The proportion of L6 neurons expressing both *Sulf1* and *Tle4* likely represents a subset of mL6-2 (Appendix C.11, A to D). *Tle4*-expressing neurons in somatosensory cortex project to the thalamus, whereas *Sulf1* is expressed by both corticothalamic and corticocortical projecting neurons [133]; hence, the projection targets of neurons in cluster mDL-2 may be different from those of neurons in clusters showing hypomethylation of *Tle4* (e.g., mL6-2). We also observed extensive overlap of in situ hybridization signals when we used probes for a classical inhibitory neuron marker gene, *Pvalb*, and *Adgra3*, a predicted mCH signature of PV inhibitory neurons (Appendix C.11, E to G), further validating the specificity of marker prediction using mCH.

We paired homologous mouse and human neuron clusters by correlating mCH levels at homologous genes and found expanded neuronal diversity in human FC relative to mouse FC (Fig. 4.2H and Appendix C.12A) [134]. Multiple human neuron clusters showed homology to mouse L5a excitatory neurons (mL5-1), L6a pyramidal neurons (mL6-2), or VIP, PV, and SST inhibitory neurons (Fig. 4.1H). We found a unique gene-specific mCH pattern and superenhancer-like mCG signatures in a potential human-specific inhibitory population (hPv-2; figs. S12B and S16J).

Although we detected substantial mCH in all human and mouse neurons, cell types varied over a wide range in terms of their genome-wide mCH level (1.3 to 3.4% in mouse, 2.8 to 6.6% in human) (Appendix C.13, A to F). The sequence context of mCH was similar across all neuron types and consistent with previous reports (Appendix C.13, I and J) [5, 4]. Interestingly, global and gene-specific mCH differences were found in PV and SST inhibitory neurons located in different cortical layers (Appendix C.14). Genes with low mCH in superficial-layer PV+ neurons are enriched in functional annotations including neurogenesis, axon guidance functions, and synaptic component (Appendix C.14, F to H), suggesting layer-specific epigenetic regulation of

synaptic functions in inhibitory neurons.

A key advantage of single-cell methylome analysis is the ability to obtain regulatory information from the vast majority of the genome (>97% [134]) not directly assessed by RNA sequencing. By pooling reads from all neurons in each cluster, we could find statistically significant differentially methylated regions with low mCG in specific neuronal populations (CG-DMRs), which are reliable markers for regulatory elements [5]. We found 575,524 mouse (498,432 human) CG-DMRs with average size of 263.6 bp (282.8 bp), covering 5.8% (5.0%) of the genome (Fig. 4.3A, Appendix C.15A, and tables S5 and S6). Most CG-DMRs (73.2% in mouse, 68.6% in human) are located >10 kb from the nearest annotated transcription start site (Appendix C.15, B to E). mPv and mVip CG-DMRs showed the strongest overlap with ATAC-seq peaks and putative enhancers identified from purified PV+ and VIP+ populations, respectively (Appendix C.15, G and H) [135]. Hierarchical clustering of mCG levels at CG-DMRs grouped neuron types by cortical layer and inhibitory neuron subtypes (Appendix C.15, I and J). Thus, neuron type classification is supported by the epigenomic state of regulatory sequences.

We inferred transcription factors (TFs) that play roles in neuron type specification by identifying enriched TF-binding DNA sequence motifs in CG-DMRs (Fig. 4.3, B and C, and Appendix C.15K). We identified known transcriptional regulators and observed that several TF-binding motifs were enriched in human but depleted in mouse CG-DMRs in homologous clusters (Fig. 4.3C). The binding motif of NUCLEAR FACTOR 1 (NF1) was enriched in CG-DMRs for two human inhibitory neuron subtypes (hVip-2, hNdnf) but was depleted in homologous mouse clusters (mVip, mNdnf-2), suggesting a specific involvement of NF1 in human inhibitory neuron specification. Thus, although the TF regulatory circuits governing tissue types are conserved between mouse and human [136], fine-grained distinctions between neuronal cell types may be shaped by species-specific TF activity.

Superenhancers are clusters of regulatory elements, marked by large domains of mediator binding and/or the enhancer histone mark H3K27ac, that control genes with cell type-specific

roles [137]. Extended regions of depleted mCG (large CG-DMRs) are also reliable markers of superenhancers (Appendix C.16, A to C) [103]. Therefore, we used our neuron type-specific methylomes to predict superenhancers for each mouse and human neuron type (Appendix C.16, D to I, and tables S7 and S8). For example, superenhancer activity was indicated by a large CG-DMRs at *Bcl11b* (*Ctip2*) in a subset of deep-layer neurons (Appendix C.16, F and G) and broad H3K27ac enrichment in mouse excitatory neurons (Appendix C.16F). Superenhancers overlap with key regulatory genes in the associated cell type, such as *Prox1* in VIP+ and NDNF+ neurons (Appendix C.16, H and I).

Global mCH and mCG levels were correlated between homologous clusters across mouse and human (Pearson $r = 0.698$ for mCH, $r = 0.803$ for mCG; $P < 0.005$), suggesting evolutionary conservation of cell type-specific regulation of mC (Fig.4.4A and Appendix C.13, G and H). Examining 12,157 orthologous gene pairs, we found stronger correlation of gene body mCH between homologous clusters in mouse and human (median Spearman $r = 0.236$; Fig. 4.4, B and C) than between different cell types within the same species ($r = -0.050$, mouse; $r = -0.068$, human). For homologous clusters, we found shared and species-specific CG-DMRs based on sequence conservation (liftover; Appendix C.17, A and B). Cross-species correlation of mCG at CG-DMRs was significantly greater for inhibitory than for excitatory neurons ($P < 0.001$, Wilcoxon rank sum test; Fig. 4.4D and Appendix C.17C). Greater sequence conservation at inhibitory neuron CG-DMRs could partly explain the greater regulatory conservation ($P < 0.001$, Wilcoxon rank sum test; Fig. 4.4E). Sequence conservation was observed only within 1 kb of the center of inhibitory neuron CG-DMRs and did not extend to the flanking regions (Appendix C.17G). These results support conservation of neuron type-specific DNA methylation, with greater conservation of inhibitory than of excitatory neuron regulatory elements.

Single-cell methylomes contain rich information enabling high-throughput neuron type classification, marker gene prediction, and identification of regulatory elements. Applying a uniform experimental and computational pipeline to mouse and human allowed unbiased comparison

of neuronal epigenomic diversity in the two species. The expanded neuronal diversity in human, revealed by DNA methylation patterns, is consistent with more complex human neurogenesis, such as the presence of outer radial glia and the potential dorsal origin of certain interneuron subtypes [130, 138, 139]. Further anatomical, physiological, and functional experiments are needed to characterize the DNA methylation-based neuronal populations defined by our study. Single-neuron epigenomic profiling allowed the identification of regulatory elements with neuron type-specific activity outside of protein-coding regions of the genome. We expect that the single-nucleus methylome approach can be applied to studies of disease, drug exposure, or cognitive experience, thereby enabling examination of the role of cell type-specific epigenomic alterations in neurological or neuropsychiatric disorders.

4.2 Acknowledgements

Chapter 4, in full, is a reprint of the material as it appears in *Science* 2017. Chongyuan Luo*, Christopher L. Keown*, Laurie Kurihara, Jingtian Zhou, Yupeng He, Junhao Li, Rosa Castanon, Jacinta Lucero, Joseph R. Nery, Justin P. Sandoval, Brian Bui, Terrence J. Sejnowski, Timothy T. Harkins, Eran A. Mukamel, M. Margarita Behrens, Joseph R. Ecker. The dissertation author was the co-investigator and co-first author of this paper.

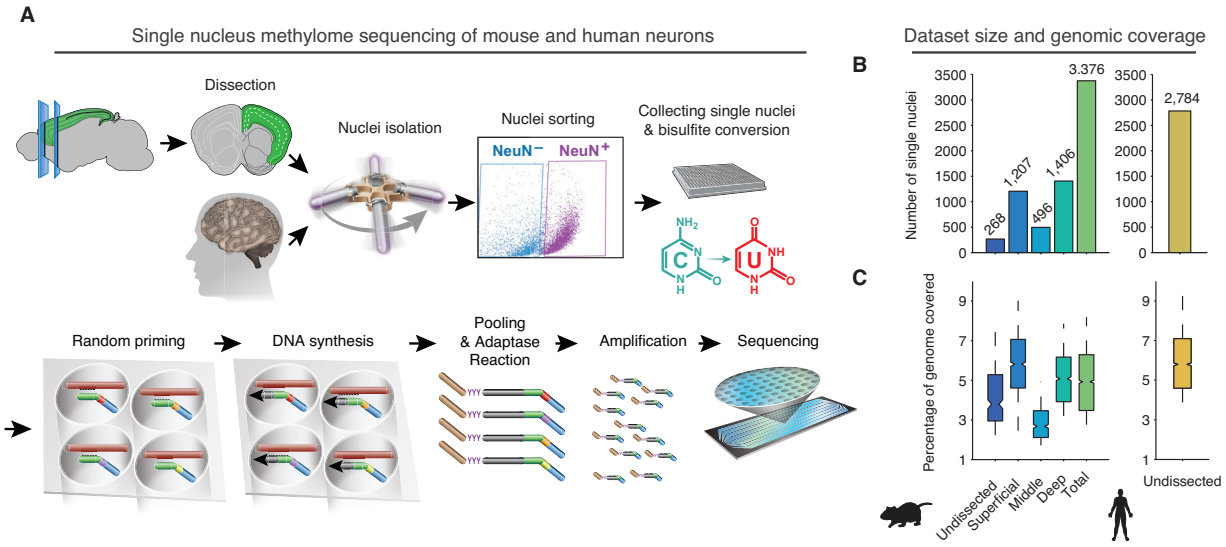

Figure 4.1: High-throughput single-nucleus methylome sequencing (snmC-seq) of mouse and human frontal cortex (FC) neurons. (**A**) Workflow of snmC-seq. (**B** and **C**) Number of single-neuron methylomes (B) and distribution of genomic coverage per data set (C).

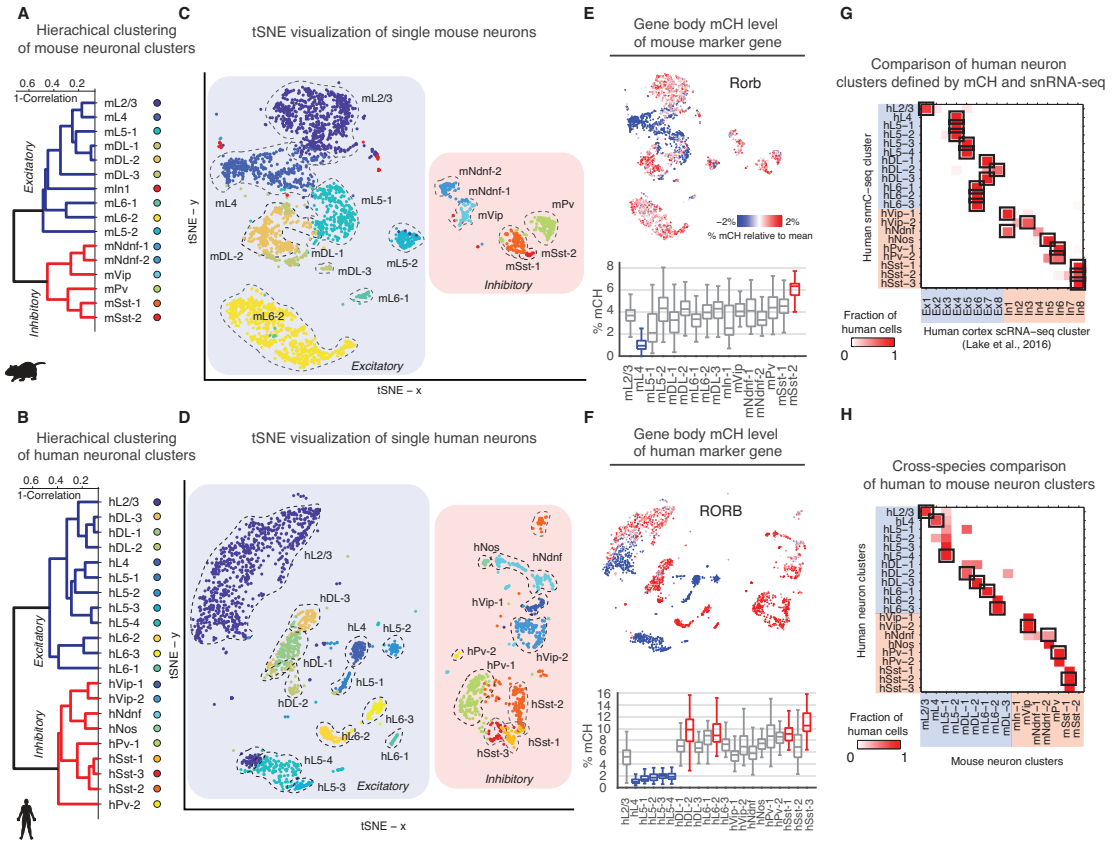

Figure 4.2: Non-CG methylation (mCH) signatures identify distinct neuron populations in mouse and human FC. (**A** and **B**) Hierarchical clustering of neuron types according to gene body mCH level. (**C** and **D**) Two-dimensional visualization of single neuron clusters (tSNE). Mouse and human homologous clusters are labeled with similar colors. (**E** and **F**) Gene body mCH at Rorb for each single neuron (top) and the distribution for each cluster (bottom); hyper- and hypomethylated clusters are highlighted in red and blue, respectively. (**G**) Comparison of human neuron clusters defined by mCH with clusters from single-nucleus RNA sequencing [39]. (**H**) Fraction of cells in each human cluster assigned to each mouse cluster based on mCH correlation at orthologous genes. Mutual best matches are highlighted with black rectangles.

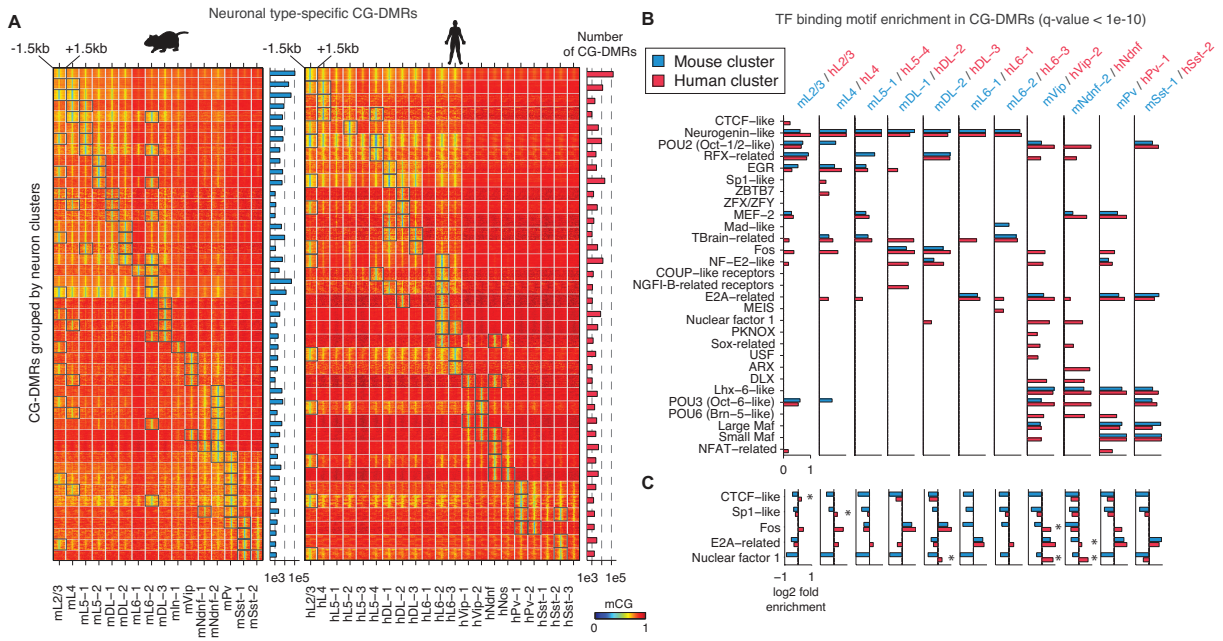

Figure 4.3: Conserved and divergent neuron type-specific gene regulatory elements. (**A**) Heat map showing differentially methylated regions (CG-DMRs) hypomethylated in one or two neuron clusters; categories of DMRs containing >1000 regions are shown. (**B**) TF binding motif enrichment in CG-DMRs of homologous mouse and human clusters (false discovery rate $< 10^{-10}$). (**C**) Mouse- or human-specific enrichment and depletion of TF binding motifs. Asterisks indicate TF binding motifs that are significantly enriched in one species but depleted in the other.

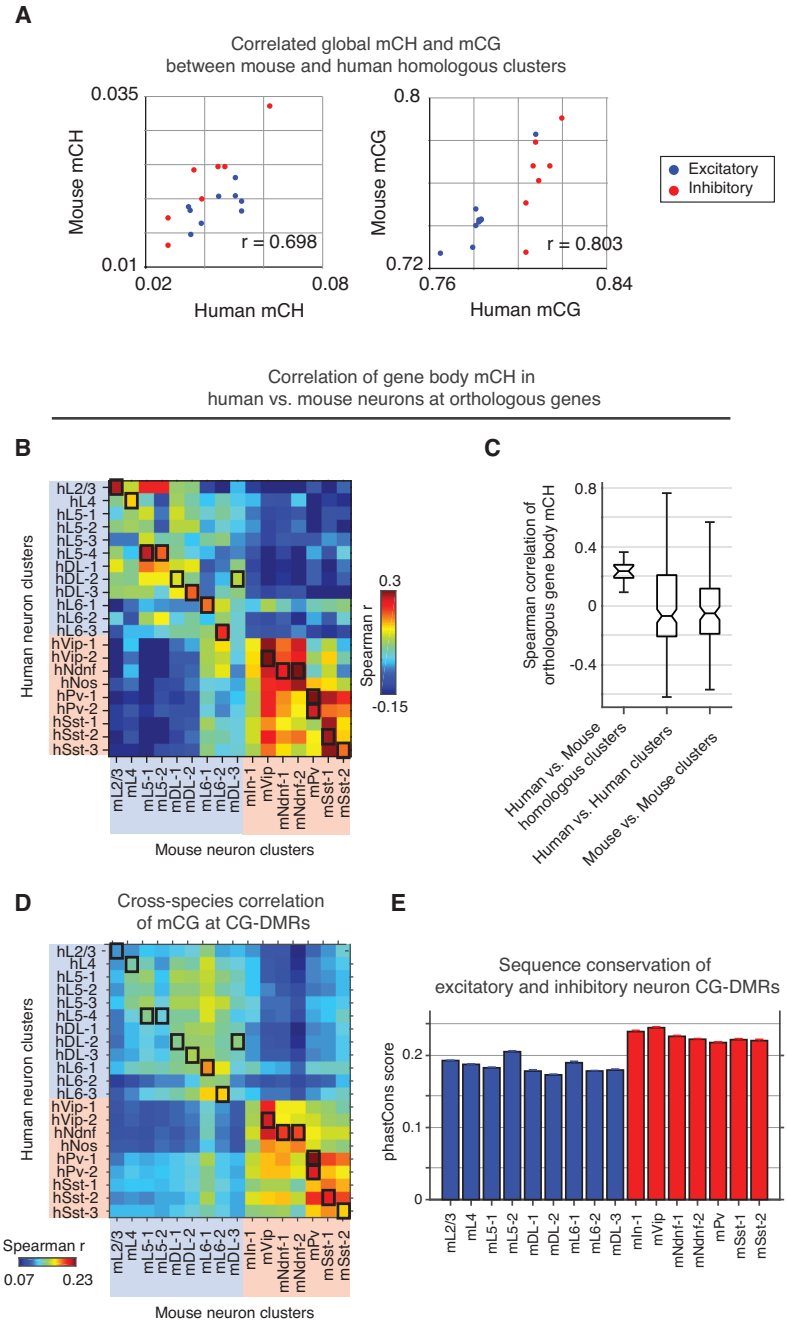

Figure 4.4: Gene body mCH and CG-DMRs conserved between mouse and human. (**A**) Global mCH and mCG levels are strongly conserved within homologous cell types between mouse and human. (**B**) Cross- species correlation of gene body mCH at orthologous genes shows cell type-specific conservation. Black boxes denote homologous neuron clusters. (**C**) The median correlation of gene body mCH for homologous clusters is higher than the within-species correlation for distinct clusters. (**D**) Cross-species correlation of mCG at neuron type-specific CG-DMRs. (**E**) Sequence conservation at neuron type-specific DMRs.

Chapter 5

Conclusions

In this dissertation, we examined the dynamics of DNA methylation in in mammalian brain cells. A key aim was to further examine the role of the recently discovered CH methylation in the brain. In the XCI study, we observed that mCH on Xi was primarily at genes escaping XCI, whereas mCH was virtually absent from the remainder of Xi. These results suggest that mCH marks active chromatin regions and that chromatin accessibility may be required for the deposition of mCH. Our examination of methylation in the dorsal and ventral regions of the DG identified a remarkable asymmetry between two structurally similar regions, with large regions of hypomethylation in the dorsal relative to ventral DG. The large methylation and transcriptional differences we identified suggest that methylation can alter anatomically similar regions and could underpin known functional differences between these regions. Finally, our analysis of single cell methylomes further support that mCH profiles are cell type specific and strongly correlated with transcriptional results from scRNA-seq studies. Furthermore, we observed large DNA methylation valleys at putative super-enhancers, supporting an association between mCH and the establishment and maintenance of cell identify. Our studies are largely correlation and a thorough examination of the functional consequences of mCH remains to be seen. As methylation assays using CRISPR or related technologies emerge, DNA methylation could be manipulated in

a systematic and targeted fashion and thereby clarify the true functional role of mCH.

The heterogeneity of tissues has long been a challenge for fully understanding molecular mechanism of neurobiological phenomenon. The development of single cell assays, including our snmC-seq and single cell RNA-seq assays, increase the resolution of our scientific lens and allow us to isolate molecular mechanisms involved in development and disease. For example, Williams syndrome emerges from the deletion of a region on chromosome 7 in humans that impacts visual-spatial and social cognition, yet the role individual genes from this genomic region play in the cognitive phenotype is poorly understood. Our data illuminates cell type specific association of these genes that may hone in on the source of these cognitive changes. By understand the role of these genes in Williams syndrome, we further understand the role of cell types in typical cognition. Single cell methylomes could be used to address limitations in our other two studies. In our examination of EE in DG, we found a small signature of methylation differences between EE and SH that overlapped with binding sites of NeuroD1, an important neurodevelopmental transcription factor. Given known differences in the underlying cell types composition of SH and EE animals in DG, at least in part from neurogenesis, future studies using single cell methylomes could identify the source of this effect and discriminate between underlying changes in cell populations and actual changes in methylation. In additional, single cell methylomes could identify which cell types within the DG underpin the dorsal and ventral differences we observe. In the X-inactivation experiment, we used whole brain tissue and our signal was thus largely dominated by excitatory neurons. Follow up single cell methylomes could provide a neuron type-specific characterization of escape genes and imprinting that could illuminate the role neuron types play in sex differences in cognition and imprinting disorders such as Prader-Willi and Angelman syndromes.

Finally, in this dissertation we generated single cell methylomes in order to characterize the neuronal heterogeneity in mouse and human. We found an expansion of deep layer cortical neurons relative to superficial layers in both mouse and human. In humans we found an increased

expansion of both excitatory and inhibitory neuron types relative to mouse. Although it is premature to fully understand the significance of these differences, our study does provide putative targets for the development of experimental models that can isolate and interrogate the role of neuron types in the context of cognition. Key questions remain unanswered about the architecture of cortex including the heterogeneity of cell of types across cortical regions, the role of genetic variability plays in neuron types across individual, and the extent individual experiences alter the molecular profile of neurons. Importantly, our snmC-seq provides an experimental technique for addressing these questions.

Appendix A

Supporting Information Chapter 2

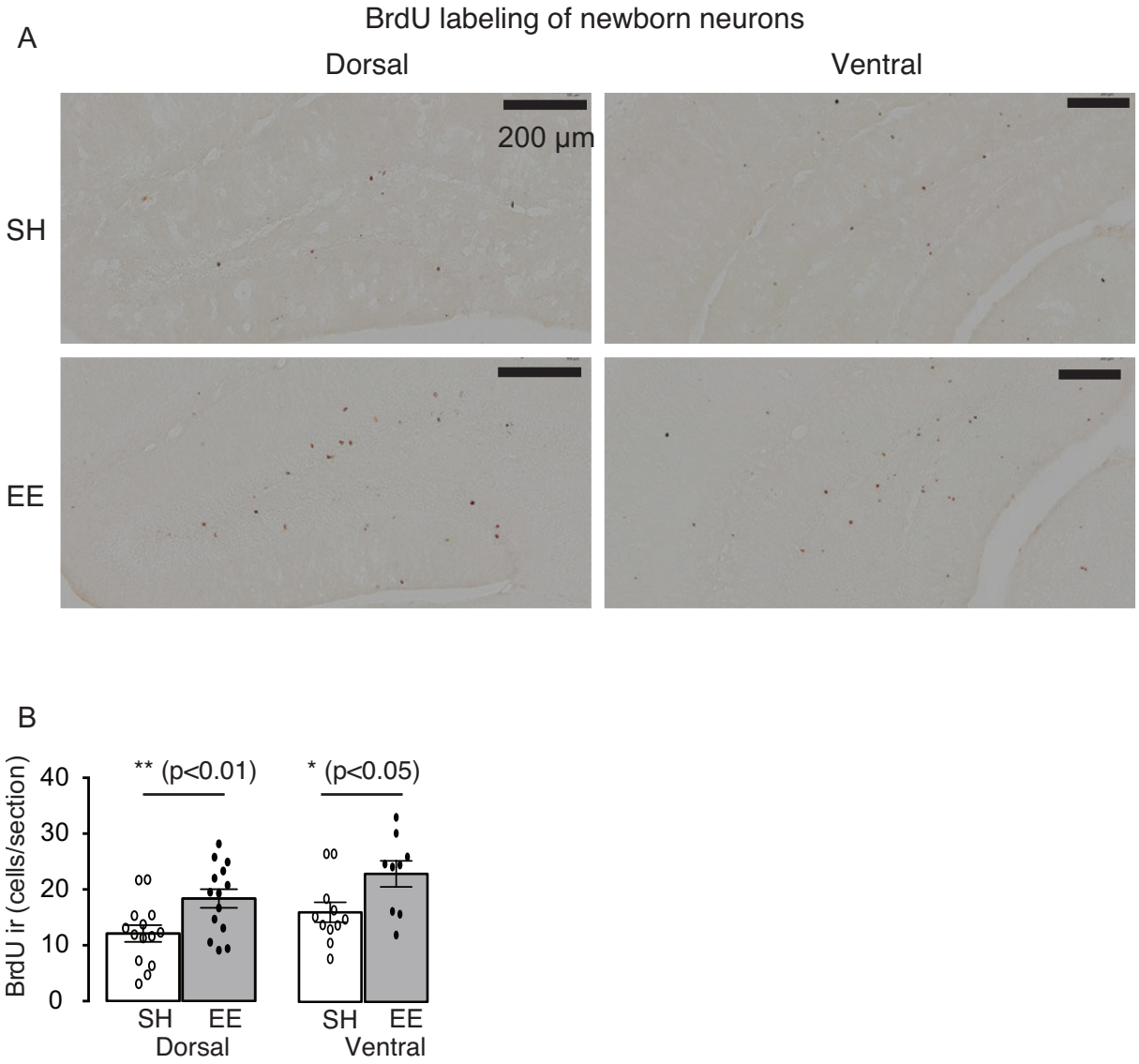

Figure A.1: Increased adult neurogenesis in dorsal and ventral DG following environmental enrichment. (A) BrdU immunoreactive cells in coronal section of dorsal and ventral dentate gyrus of mice. (B) Mean ± SEM BrdU-positive cells in dorsal and ventral dentate gyrus of mice (n = 9-14 per group, F (1,44) = 13.33, $p = 0.0007$).

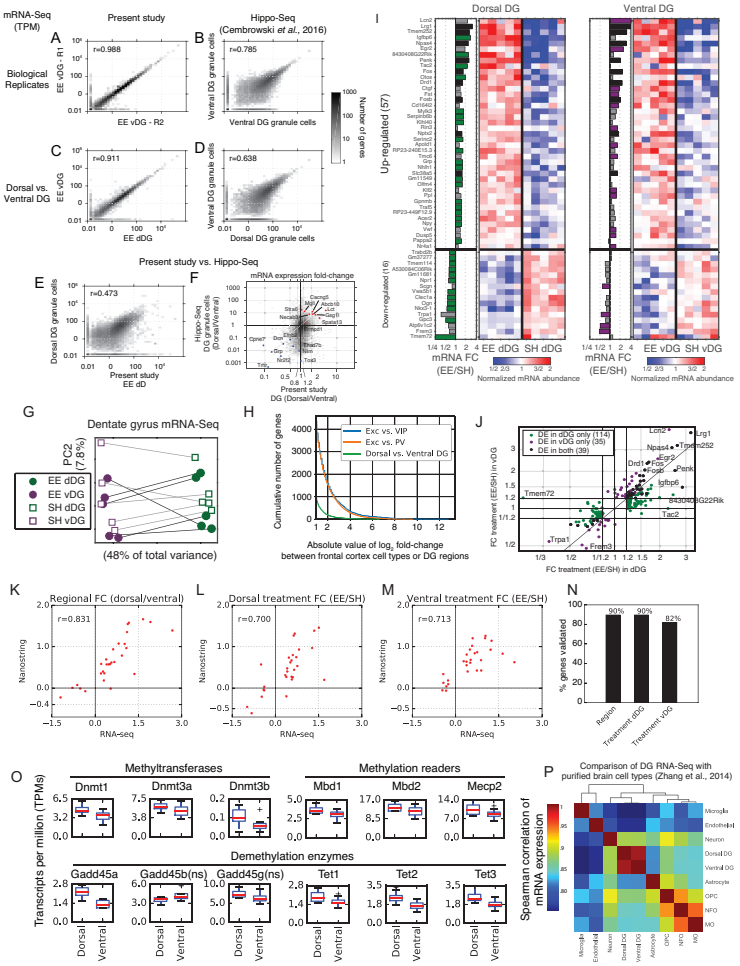

Figure A.2: Transcriptome analysis of dorsal and ventral DG in enriched environment. (**A**,**B**) Comparison of biological replicates shows the consistency of gene expression estimates (TPM) in ventral DG in the present study (A) and for ventral DG granule cells in Hippo-Seq2 (B). (**C**,**D**) Dorsal vs. ventral expression estimates for the present study (C) and hippo-seq (D). (**E**) Direct comparison of individual replicates from the present study with hippo-seq granule cells. (**F**) Comparison of regional differences in expression between the present data set (DG tissue) and purified granule cells from Hippo-Seq [40] shows highly consistent differential expression for markers of dorsal and ventral DG granule cells. (**G**) Effects of EE on gene expression are larger in dDG compared with vDG. (**H**) Dorsal and ventral DG expression differences are ∼4-fold smaller than the differences between distinct cortical cell types. (**I**) The top DE genes include immediate early genes (*Fos*, *Egr2*, *Npas4*). (**J**) More genes are upregulated (EE¿SH) than downregulated (EE¡SH). (**K-N**) Nanostring digital quantification validates RNA-Seq results. (**O**) Expression of genes associated with (de)methylation and methylation readers in dorsal and ventral DG. All genes are significantly upregulated in dorsal DG over ventral (FDR ¡ .05) except *Gadd45b*, which is significantly upregulated in ventral DG, and *Gadd45g*, which is not differentially expressed. Not significant (ns). (**P**) Correlation of expression levels between DG tissue and purified brain cell types [52] reveals expression in DG is most strongly correlated with neurons. *Oligodendrocyte progenitor cells (OPC), newly formed oligodendrocytes (NFO), mature oligodendrocytes (MO).*

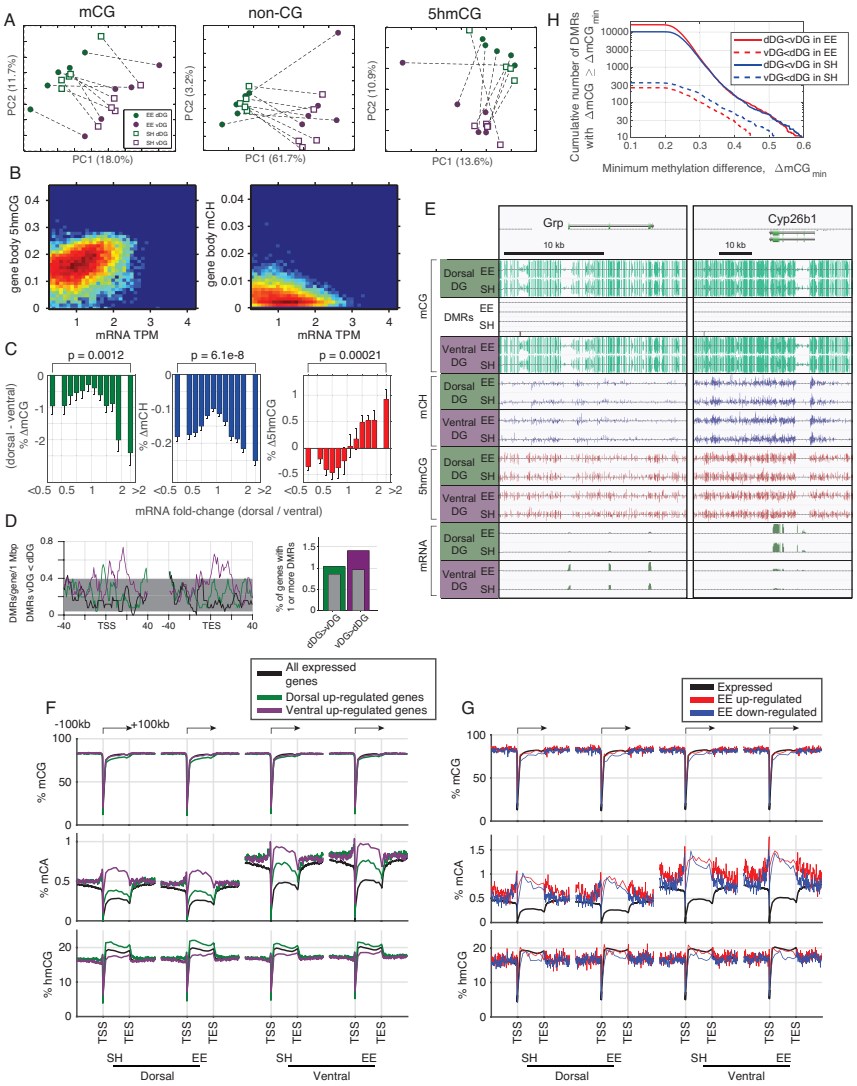

Figure A.3: (**A**) Transcriptome principal components separate dorsal and ventral samples using mCG, mCH or 5hmCG. Dashed lines connect dorsal and ventral DG from the same group of mice. (**B**) Gene body mCH is associated with transcriptional repression, whereas 5hmCG is positively correlated with expression [4, 60]. (**C**) Genes up-regulated in dorsal vs. ventral DG have lower promoter mCG, gene body mCH, and higher gene body 5hmCG compared with down-regulated genes. (**D**) Ventral hypo-methylated DMRs are enriched within the gene body of genes up-regulated in ventral DG. (**E**) Browser shots showing examples of genes up-regulated in ventral (*Grp*) or dorsal (*Cyp26b1*) DG, but which lack DMRs. (**F**,**G**) Mean mCG, mCA and hmCG levels within gene bodies $\pm$100 kb are shown for all expressed genes, as well as genes differentially expressed in dorsal or ventral DG. mCA is depleted, and hmCG is enriched, within the gene bodies of dorsal expressed genes (green lines); the opposite pattern prevails within the bodies of ventral expressed genes (purple lines). (**H**) Number of hypomethylated DMRs in dorsal (solid lines) and ventral (dashed lines) DG as a function of the methylation difference threshold.

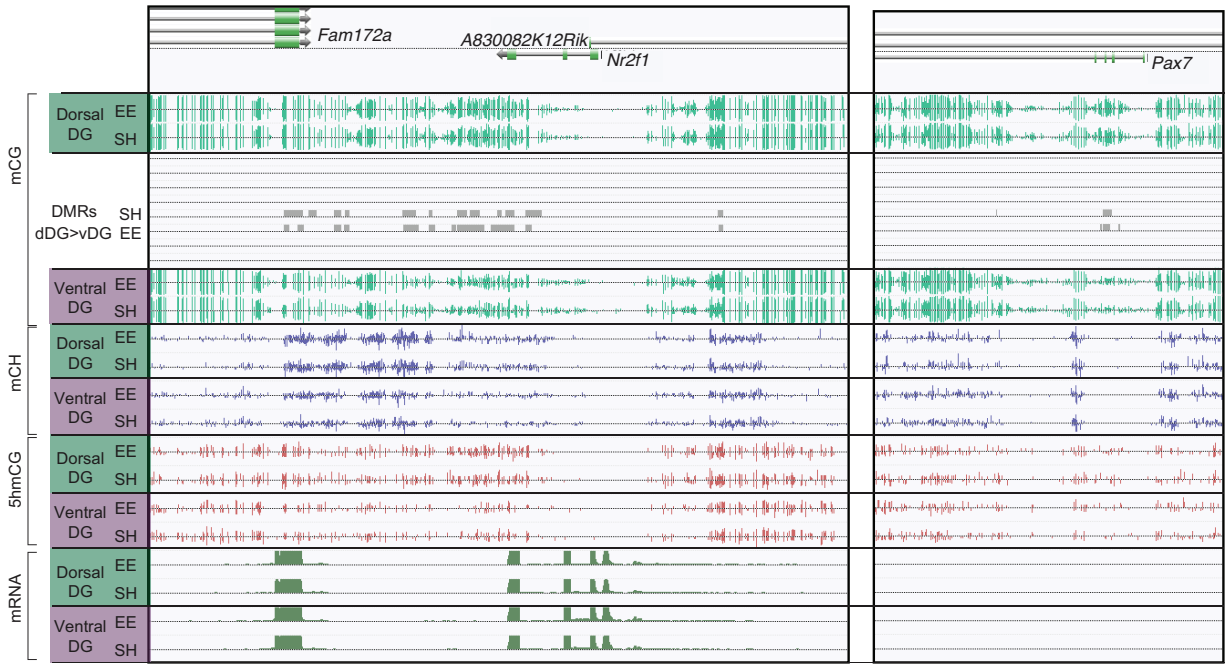

Figure A.4: Browser views of examples of genes hypomethylated in ventral DG. (Left) *Nr2f1* is differentially expressed in adult ventral DG, consistent with the lower mCG level in the gene body and surrounding region. (Right) *Pax7* is not expressed in the adult DG, but is expressed during development and helps to define dorsal fate [71]. Interestingly, we find that *Pax7* is more methylated in the dorsal DG, which may be a vestigial signature of its early developmental activity in the dorsal DG similar to vestigial DMRs observed in other neuronal cell types [5].

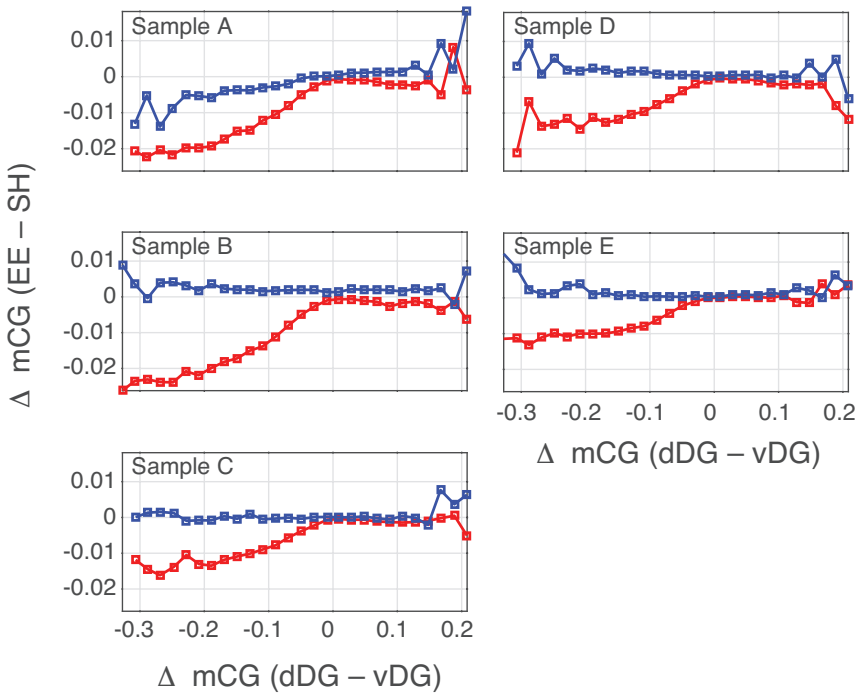

Figure A.5: Cross-validated analysis of the correlation of region- and treatment-based effects on DNA methylation. Each panel shows the result of analyzing regional DNA methylation in one of the five biological replicates (samples A-E, x-axis) vs. the treatment-based effects estimated using the remaining four samples. DNA methylation was estimated within 1kb bins across all autosomes; we excluded bins with < 10 CG basecalls in any sample, or < 100 basecalls after summing all samples. Binned mCG levels were used to compute the dorsal-ventral and EE-SH difference in methylation for each bin. We then grouped all bins according to their level of dorsal-ventral difference in mCG, and computed the median EE-SH difference.

Appendix B

Supporting Information Chapter 3

B.1 SI Discussion: Validation of Allele-Specific Methylation Accuracy

Our allele-specific analyses are based on assignment of reads containing one or more SNPs differing between C57 and Spretus reference genomes. This threshold was chosen to maximize genomic coverage, and resulted in $> 70\%$ coverage of each allele. Here, we describe two analyses that we performed to validate the accuracy of our allele-specific computational pipeline. First, we reanalyzed our data using a more stringent criterion for allele assignment: We required at least two SNPs per read, instead of one SNP per read as used in our paper. As expected, because many reads cover just one SNP, this requirement resulted in a lower rate of allele assignment (33.5% instead of 68.6% using all reads with one or more SNPs). Nevertheless, we found that our results were quantitatively unchanged when using this more conservative analysis, with high Pearson correlations between the two analyses for gene body mCH on Xa ($r = 0.981$) and Xi ($r = 0.986$), as well as for promoter mCG on Xa ($r = 0.996$) and Xi ($r = 0.990$).

A second line of reasoning allows us to provide a stringent bound on the possible rate of allele misassignment for our pipeline. Any misclassification of reads (i.e., incorrect assignment

of C57 reads to Spretus or vice versa) should lead to apparently more similar methylation levels on the two alleles. In particular, within regions of the X chromosome that lack escape genes, we found apparent mCH levels of 0.93% on Xa and 0.017% on Xi (after correcting for the bisulfite nonconversion rate). The maximum possible rate of misclassification consistent with these observations would occur if the true methylation levels were 0% (Xi) and 0.947% (Xa), which would correspond to a false assignment rate of 0.017/0.947 = 1.8%. Even allowing for a possible $\sim$ 10% overestimation of the nonconversion rate, we still find that the false allele assignments are $\leq$ 5% in the worst case. The true false assignment rate is likely lower than this bound because of the robustness of our results when we used a more stringent allele-assignment procedure (discussed above).

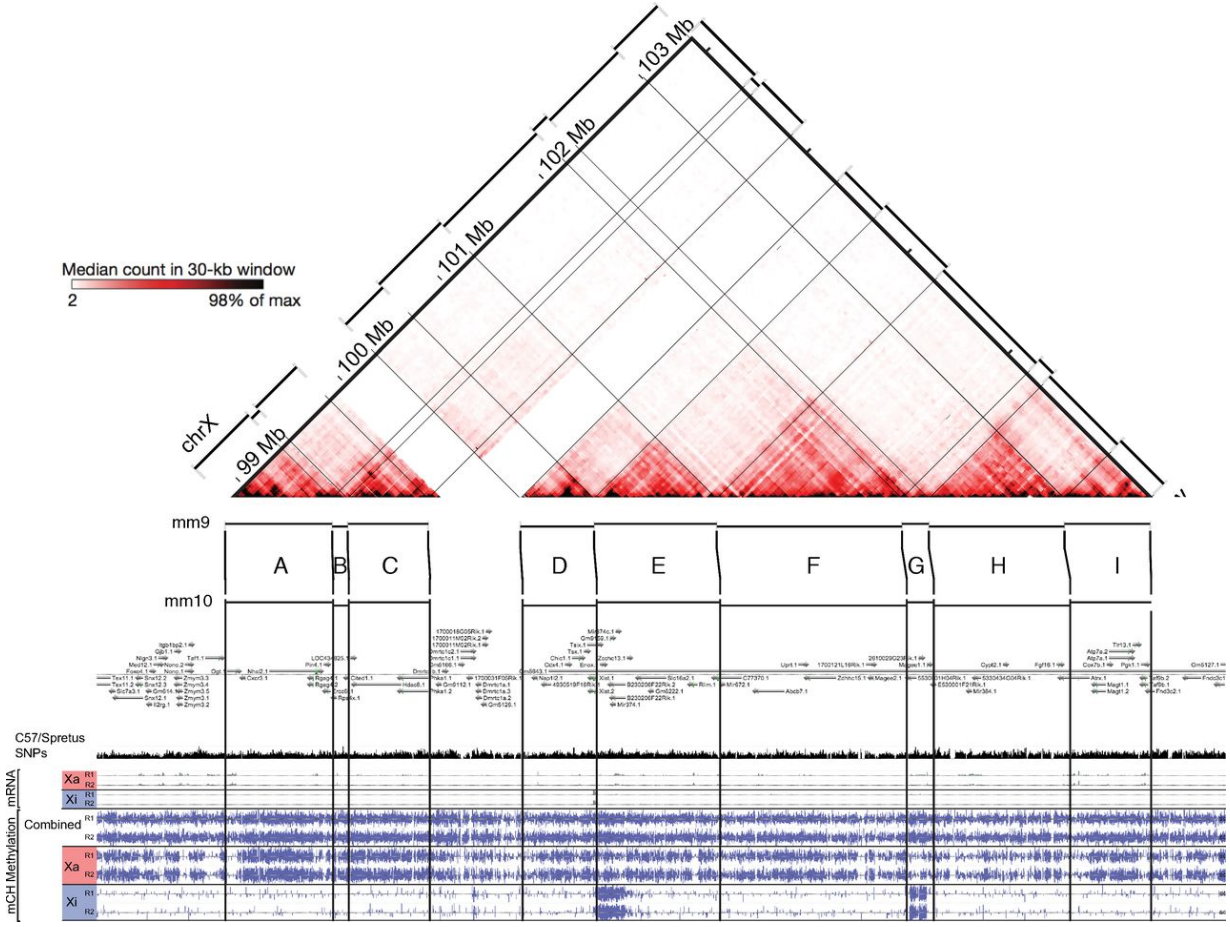

Figure B.1: Topologically associated domains correlate with mCH. Coordinates for the TADs on Xi at the XCI [105] were converted to mm10 using liftOver [140] and superimposed on a browser view of mCH. Boundaries between domains are shown as black lines and overlap with regions of hyper-mCH on Xi, particularly in TADs E and G.

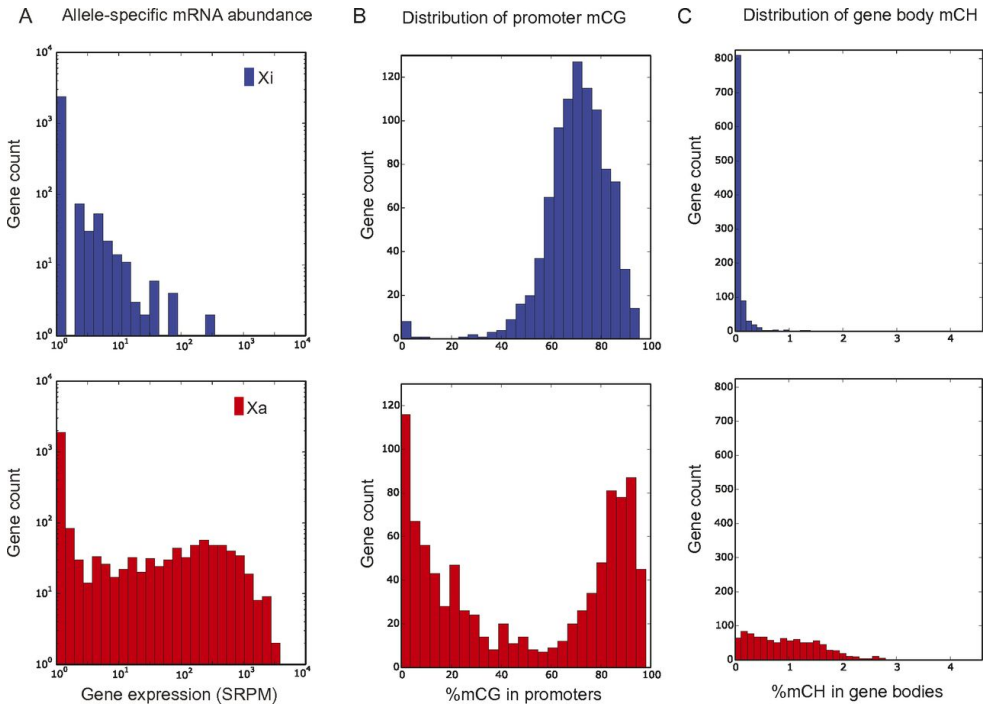

Figure B.2: Asymmetrical expression and methylation distinguish Xa from Xi. Histograms of gene expression on Xa and Xi in SRPMs (**A**), CG methylation in promoters (**B**), and CH methylation in gene bodies (**C**) are shown.

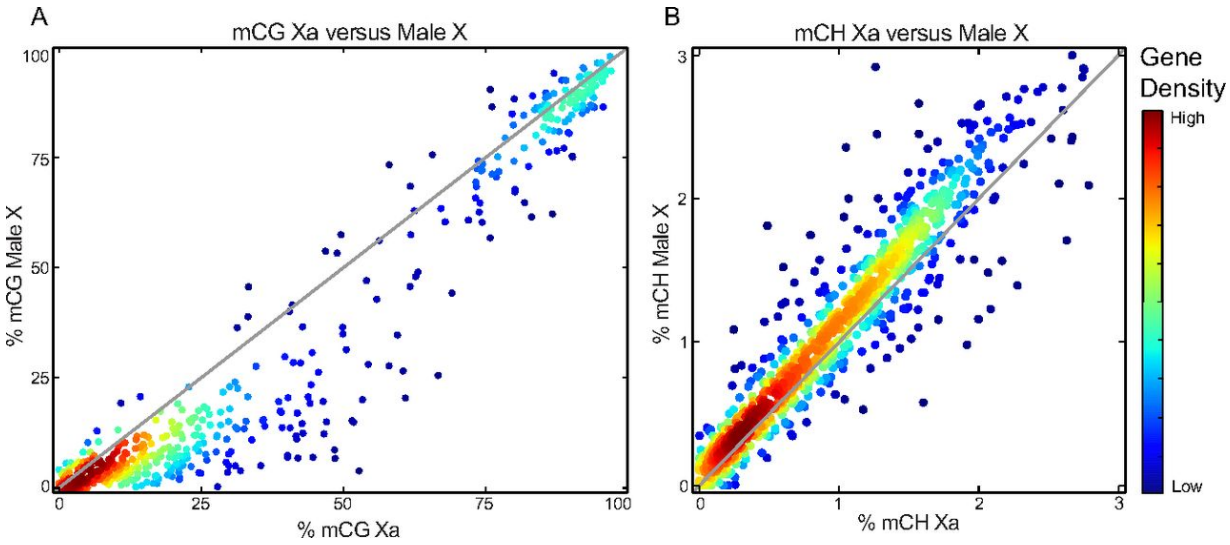

Figure B.3: Methylation on Xa and the male X chromosome are similar. Methylation levels on Xa and the male X chromosome for mCG in promoters (**A**, $r = 0.96$) and mCH in gene bodies (**B**, $r = 0.94$). Points correspond to individual genes, and the point color indicates the density of genes in that region.

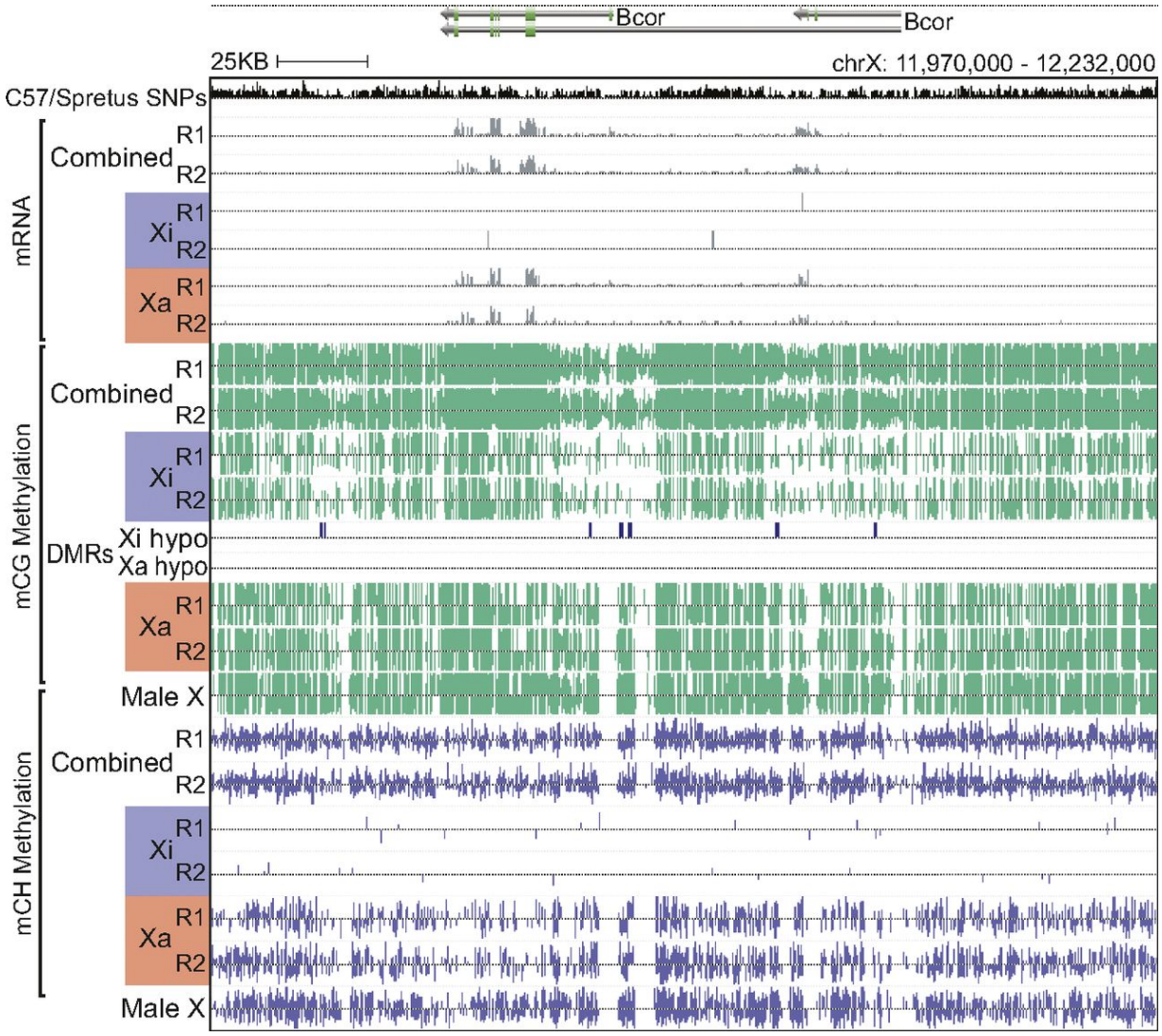

Figure B.4: *Bcor* shows diffuse CG hypomethylation on Xi, although it does not escape. A browser view shows methylation and expression at the gene *Bcor*.

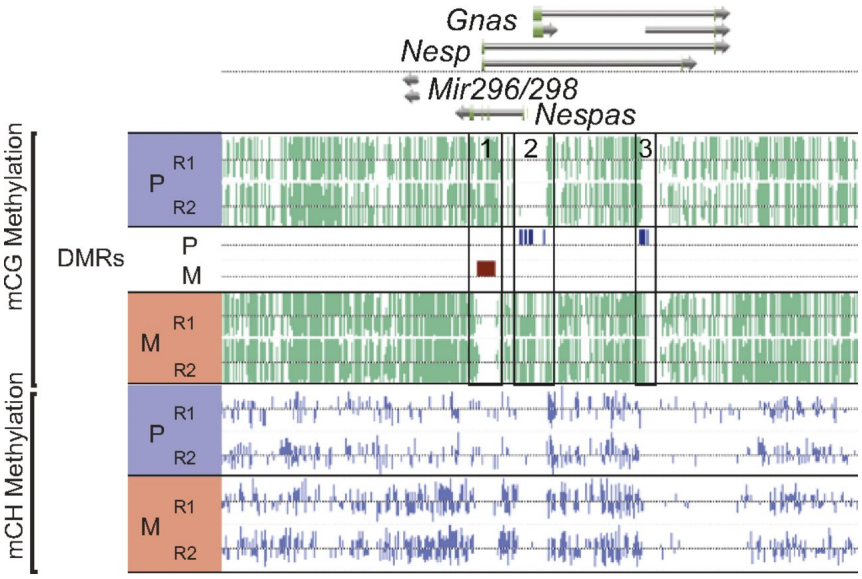

Figure B.5: Bidirectional allele-specific methylation at the *Nesp*/*Gnas*/*Nespas* locus. A browser view of mCG and mCH shows three neighboring DMRs with opposite polarity. The promoter of *Nesp* (1) lacks mCG on the maternal allele, whereas the promoters of *Gnas* (3) and the antisense transcript *Nespas* (2) lack mCG on the paternal allele. These CG-DMRs are consistent with reported maternal and paternal imprinted expression of *Nesp* and *Nespas*, respectively [141]. Overall, this region demonstrates lower mCH on the paternal allele, showing a consistent asymmetry with the *Nesp* gene but the reverse asymmetry compared with *Nespas*. M, maternal; P, paternal.

Appendix C

Supporting Information Chapter 3

C.1 Materials and Methods

C.1.1 Animal samples

For the production of single neuron methylomes from layer dissected mouse frontal cortex tissue, eight week old C57BL/6J male mice were purchased from Jackson Laboratories, Bar Harbor ME, and allowed a week of acclimation in our animal facility with 12 h light/dark cycles and food ad libitum before sacrificing and dissecting.

Nuclei were also isolated from frontal cortex of the CLSun1-G35-Cre line with no layer dissection. This line was produced by crossing the transgenic line B6;129-Gt(ROSA)26Sortm5 (CAG-Sun1/sfGFP)Nat/J (described in [5], but backcrossed into a C57BL/6J background for 9 generations), with the G35-Cre line [142].

Nuclei of SST+ inhibitory neuron population was isolated from frontal cortex of CLSun1-SST-Cre line. This line was generated by crossing B6;129-Gt(ROSA)26Sortm5(CAG-Sun1/sfGFP) Nat/J backcrossed into C57BL/6J with SST-Cre line (Jackson Labs).

All protocols were approved by the Salk Institute's Institutional Animal Care and Use Committee (IACUC).

C.1.2 Human samples

The human brain specimen was obtained from the NICHD Brain and Tissue Bank for Developmental Disorders at the University of Maryland, Baltimore, MD. The frozen middle frontal gyrus tissue belonged to a 25-year-old Caucasian male (UMB#4540) with a PMI of 23 h.

C.1.3 Mouse tissue dissections

To produce the frontal cortex tissue, mouse brains were sectioned coronally at Bregma 2.5 and 0.5 with a razor blade in dissection media [20 mM Sucrose, 28 mM D-Glucose (Dextrose), 0.42 mM NaHCO3, in HBSS]. For cortical layer dissection, the tissue block (devoid of non-cortical tissue) was then dissected under a microscope (SZX16, Olympus). The cortical region was divided parallel to the meninges into three sections of approximately equal width such that "superficial layers" contained layers 1-3 with part of layer 4, and "deep layers" contained mainly layers 6 and part of 5.

C.1.4 Nuclear isolation

Nuclear isolation from mouse and human cortical tissues was performed as described in [4] with the following modifications: Proteinase inhibitor (11836153001, Roche) and RNAse inhibitor (30 U/ml, PRN2611 from Promega) were added to the lysis buffer and sucrose gradients. After centrifugation, nuclei were resuspended in 0.5% BSA (AM2616, Ambion) and PBS (Ca_{2+} and Mg_{2+} free, 14190-144 from Life Technologies) with protein and RNAse inhibitors.

C.1.5 Flow cytometry based nuclei sorting

Isolated nuclei from mouse and human tissues were labeled by incubation with 1:1000 dilution of AlexaFluor488 conjugated anti-NeuN antibody (MAB377X, Millipore) at 4°C for 1 hour. Nuclei isolated from CLSun1-G35-Cre line were incubated with AlexaFluor647 conju-

gated anti-NeuN antibody (anti-NeuN antibody MAB377 labeled using Apex Alexa Fluor 647. A10475, Life Technologies) and AlexaFluor488 conjugated anti-GFP antibody (A21311, Life Technologies). Fluorescence-activated nuclei sorting (FANS) of single nuclei was performed using a BD Influx sorter with an 85 μm nozzle at 22.5 PSI sheath pressure. Single nuclei were sorted into each well of a 384-well plate preloaded with 2 μl of Proteinase K digestion buffer (1 μl M-Digestion Buffer, 0.1 μl 20 μg/μl Proteinase K and 0.9 μl H_2O). The alignment of the receiving 384-well plate was performed by sorting sheath flow into wells of an empty plate and making adjustments based on the liquid drop position. Single cell (1 drop single) mode was selected to ensure the stringency of sorting.

C.1.6 Preparation of single nucleus methylome library

Steps of library preparation prior to SPRI purification were performed in a horizontal laminar flow hood to minimize environmental DNA contamination. Bisulfite conversion of single nuclei was carried out using Zymo EZ-96 DNA Methylation-DirectTM Kit (Deep Well Format, cat. #D5023) following the product manual with reduced reaction volume. 384-well plates (ThermoFisher Armadillo PCR Plate cat. # AB2384) containing FACS isolated single nuclei were heated at 50°C for 20 min. 25 μl CT Conversion Reagent was added to each well, followed by pipetting up and down to mix. Plates were treated with the following program using a thermocycler: 98°C for 8 min, 64°C for 3.5 hours and 4°C forever.

Each well of Zymo-Spin I-96 Binding Plates was preloaded with 150 μl M-binding buffer. Bisulfite conversion reactions were transferred from 384-well plates to I-96 Binding Plates followed by pipetting up and down to mix. I-96 Binding Plates were centrifuged at 5,000g for 5 min. Wells were washed with 400 μl of M-Wash Buffer, followed by centrifugation at 5,000g for 5 min. 200 μl of M-Desulphonation Buffer were added to each well and incubated for 15 min at room temperature before removed by centrifugation at 5,000g for 5 min. Each well was then washed with 400 μl of M-Wash Buffer twice. 12 μl of M-Elution Buffer were added to each well

and incubated for 5 min at room temperature. I-96 Binding Plate was placed above a 96-well PCR plate (Applied Biosystems MicroAmp EnduraPlateTM cat. # 4483348) and was centrifuged at 5,000g for 3 min. 9 μl of eluted DNA were commonly collected in each well of the PCR plate.

Each of the four indexed random primers (P5L-AD002-N9, P5L-AD006-N9, P5L-AD008-N9 and P5L-AD010-N9) was used for indexing a 96-well plate containing bisulfite converted single nuclei. The four plates would be combined during a later SPRI step. 1 μl of 5 μM indexed random primer was added to each well of 96-well plate, followed by mixing with vortexing. All DNA oligos were purchased from Integrated DNA Technologies (IDT).

```
P5L-AD002-N9
/5SpC3/TTCCCTACACGACGCTCTTCCGATCTCGATGT(N1:25252525)
(N1)(N1) (N1)(N1)(N1) (N1)(N1)(N1)
P5L-AD006-N9
/5SpC3/TTCCCTACACGACGCTCTTCCGATCTGCCAAT(N1:25252525)
(N1)(N1) (N1)(N1)(N1) (N1)(N1)(N1)
P5L-AD008-N9
/5SpC3/TTCCCTACACGACGCTCTTCCGATCTACTTGA(N1:25252525)
(N1)(N1) (N1)(N1)(N1) (N1)(N1)(N1)
P5L-AD010-N9
/5SpC3/TTCCCTACACGACGCTCTTCCGATCTTAGCTT(N1:25252525)
(N1)(N1) (N1)(N1)(N1) (N1)(N1)(N1)
```

96-well plates were heated at 95°C using a thermocycler for 3 min to denature sample and were immediately chilled on ice for 2 min. 10 μl enzyme mix containing 2 μl of Blue Buffer (Enzymatics cat. # B0110), 1 μl of 10mM dNTP (NEB cat. # N0447L), 1 μl of Klenow exo- (50U/μl, Enzymatics cat. # P7010-HC-L) and 6 μl H2O, was added to well. After mixing with vortexing, plate was treated with the following program using a thermocycler: 4°C for 5 min, ramp up to 25°C at 0.1°C/sec, 25°C for 5 min, ramp up to 37°C at 0.1°C/sec, 37°C for 60 min, 4°C. 2 μl of Exonuclease 1 (20U/μl, Enzymatics cat. # X8010L) were added to each well, followed by mixing with vortexing. Plate was incubated at 37°C for 30 min and then 4°C forever using a thermocycler. μ 17.6 μl of home-made SPRI beads were added to each well. Sample/bead mixture from four plates, indexed using distinct indexed random primers, was combined and followed by

pipetting up and down to mix. Sample/bead mixture was incubated at room temperature for 5 min before being placed on a 96-well magnetic separator (DynaMag-96 Side Magnet, ThermoFisher Cat. # 12331D. DynaMag-96 Side Skirted Magnet, ThermoFisher Cat. # 12027). Supernatant was removed from each well, followed by three rounds of washing with 180 μl of 80% ethanol. After air drying beads at room temperature, 10 μl M-Elution buffer were added to each well to fully resuspend the beads. Eluted samples were transferred to a new 96-well PCR plate.

PCR plate was heated at 95°C for 3 min using a thermocycler to denature sample and was immediately chilled on ice for more than 2 min. 10.5 ul Adaptase master mix (2 μl Buffer G1, 2 μl Regent G2, 1.25 μl Reagent G3, 0.5 μl Enzyme G4, 0.5 μl Enzyme G5 and 4.25 μl M-Elution buffer; Accel-NGS Adaptase Module for Single Cell Methyl-Seq Library Preparation, Swift Biosciences, cat. # 33096) was added into each well, followed by mixing with vortexing. Plates were incubated at 37°C at 30 min and then 4°C using a thermocycler. 30 μl PCR mix (25 μl KAPA HiFi HotStart ReadyMix, KAPA BIOSYSTEMS, cat. # KK2602, 1 μl 30 μM P5 indexing primer and 5 μl 10 μM P7 indexing primer) were added into each well, followed by mixing with vortexing.

P5 Indexing primers:

```
P5L_D501
AATGATACGGCGACCACCGAGATCTACACTATAGCCTACACTCTTTCCCTACACGACGCTCT
P5L_D502
AATGATACGGCGACCACCGAGATCTACACATAGAGGCACACTCTTTCCCTACACGACGCTCT
P5L_D503
AATGATACGGCGACCACCGAGATCTACACCCTATCCTACACTCTTTCCCTACACGACGCTCT
P5L_D504
AATGATACGGCGACCACCGAGATCTACACGGCTCTGAACACTCTTTCCCTACACGACGCTCT
P5L_D505
AATGATACGGCGACCACCGAGATCTACACAGGCGAAGACACTCTTTCCCTACACGACGCTCT
P5L_D506
AATGATACGGCGACCACCGAGATCTACACTAATCTTAACACTCTTTCCCTACACGACGCTCT
P5L_D507
AATGATACGGCGACCACCGAGATCTACACCAGGACGTACACTCTTTCCCTACACGACGCTCT
P5L_D508
AATGATACGGCGACCACCGAGATCTACACGTACTGACACACTCTTTCCCTACACGACGCTCT
```

P7 indexing primers:

```
P7L_D701
CAAGCAGAAGACGGCATACGAGATCGAGTAATGTGACTGGAGTTCAGACGTGTGCTCTT
P7L_D702
CAAGCAGAAGACGGCATACGAGATTCTCCGGAGTGACTGGAGTTCAGACGTGTGCTCTT
P7L_D703
CAAGCAGAAGACGGCATACGAGATAATGAGCGGTGACTGGAGTTCAGACGTGTGCTCTT
P7L_D704
CAAGCAGAAGACGGCATACGAGATGGAATCTCGTGACTGGAGTTCAGACGTGTGCTCTT
P7L_D705
CAAGCAGAAGACGGCATACGAGATTTCTGAATGTGACTGGAGTTCAGACGTGTGCTCTT
P7L_D706
CAAGCAGAAGACGGCATACGAGATACGAATTCGTGACTGGAGTTCAGACGTGTGCTCTT
P7L_D707
CAAGCAGAAGACGGCATACGAGATAGCTTCAGGTGACTGGAGTTCAGACGTGTGCTCTT
P7L_D708
CAAGCAGAAGACGGCATACGAGATGCGCATTAGTGACTGGAGTTCAGACGTGTGCTCTT
P7L_D709
CAAGCAGAAGACGGCATACGAGATCATAGCCGGTGACTGGAGTTCAGACGTGTGCTCTT
P7L_D710
CAAGCAGAAGACGGCATACGAGATTTCGCGGAGTGACTGGAGTTCAGACGTGTGCTCTT
P7L_D711
CAAGCAGAAGACGGCATACGAGATGCGCGAGAGTGACTGGAGTTCAGACGTGTGCTCTT
P7L_D712
CAAGCAGAAGACGGCATACGAGATCTATCGCTGTGACTGGAGTTCAGACGTGTGCTCTT
```

PCR plate was treated with the following program using a thermocycler: 95°C for 2 min, 98°C for 30 sec, 17 cycles of (98°C for 15 sec, 64°C for 30 sec, 72°C for 2min), 72°C for 5 min and then 4°C. PCR products were cleaned up using 0.8x SPRI beads and were combined into one tube for each 96-well plate. Pooled PCR product was resolved on 2% agarose gel, smear between 400 bp and 2 Kb were excised and purified using QIAquick Gel Extraction Kit (Qiagen cat. # 28706). Library concentration was determined using Qubit dsDNA HS (High Sensitivity) Assay Kit (Invitrogen cat. # Q32851).

C.1.7 Sequencing of single nucleus methylome library

Pooled library concentration was adjusted to 700 - 800 pM for cluster generation and was sequenced on Illumina HiSeq 4000 instrument using RTA 2.7.7.

C.1.8 Single cell methylome mapping and data analysis

Sequencing reads were first trimmed to remove sequencing adaptors using Cutadapt 1.11 [143] with the following parameters in paired-end mode: -f fastq -q 20 -m 62 -a AGATCGGAAGAGCACACGTCTGAAC -A AGATCGGAAGAGCGTCGTGTAGGGA. For singleplex samples, -m parameter was set to 40. For multiplexed samples, 16 bp were further trimmed from both 5'- and 3'- ends of R1 and R2 reads to remove random primer index sequence and C/T tail introduced by Adaptase, with the following parameters: -f fastq -u 16 -u -16 -m 30. Trimmed reads for mouse and human single nuclei were mapped to mm10 and hg19 reference genomes, respectively. R1 and R2 reads were mapped separately as single-end reads using Bismark v0.15.0 with parameter –bowtie2. –pbat option was activated for mapping R1 reads [144]. Resulting bam files were sorted using SAMtools 1.3 sort [145], followed by removal of duplicate reads using Picard 1.141 MarkDuplicates with the option REMOVE_DUPLICATES=true (https://broadinstitute.github.io/picard/). Non-clonal reads were further filtered for minimal mapping quality (MAPQ $\geq$ 30) using samtools view with option -q30. To prevent regional mCH level estimation from being skewed by rare reads that failed to be bisulfite converted, reads with read-level mCH level greater than 0.7 were excluded.

The calling of unmethylated and methylated base calls was performed by call_methylated_sites of Methylpy (https://github.com/yupenghe/methylpy/) [4, 5, 103].

C.1.9 MethylC-seq of mouse SST+ inhibitory neurons

MethylC-seq library was constructed following the protocol described in detail in [146].

C.1.10 Genomic sequencing of the human sample

Genomic DNA was extracted from the human specimen using DNeasy Blood & Tissue Kit (Qiagen cat. # 69504). Genomic sequencing library was constructed using the same procedure as MethylC-seq library except bisulfite conversion was not performed.

C.1.11 Calling sequence variants for the human sample

Adaptor sequence was trimmed from sequencing reads using Cutadapt 1.11 with the following options: -f fastq -q 20 -m 50 -a AGATCGGAAGAGCACACGTCTGAAC -A AGATCGGAAGAGCGTCGTGTAGGGA. Trimmed reads were mapped to human hg19 reference genome using Bowtie2 2.2.5 with option -X 2000. Mapped reads were filtered for minimal mapping quality (MAPQ $\geq$ 20) using samtools view with option -q20. For calling sequence variants, a filtered bam file was processed using samtools mpileup with option -ug with the outputs piped into bcftools called with option -vm0 v [147].

C.1.12 Data cleaning

Data were cleaned by excluding low-quality cells using the following set of conservative criteria, ultimately yielding 3376 cells in mouse and 2784 cells in human for analysis. First, non-conversion rate was required to be low ($\leq$1% in mouse and $\leq$2% in human). We set a minimum on the number of non-clonal mapped reads to eliminate contaminated samples (400K in mouse; 500K in human). We also set an upper limit on coverage to protect against wells with multiple cells ($\leq$15% of cytosines).

In order to minimize contamination in the human data from exogenous human DNA fragments, we identified potentially contaminated samples using a genomic sequence variant (SNP) matching process. Single nucleotide variants identified from genomic sequencing (g) of the human sample were compared to variants observed in each single human nucleus methylome

(*m*), with hg19 serving as the reference genome for mapping both data types. SNP compatibility between *g* and *m* was scored for all homozygous variant sites identified in *g* that were covered by methylome reads in *m*. A compatible site between *g* and *m* required identical genotype between *g* and *m* at all sites where the variant sequence was A or T in *g*. For a site with variant sequence C in *g*, sequence = C or T in *m* was considered compatible. For a site with variant sequence G in *g*, sequence = G or A in *m* was considered compatible. For each single human nucleus methylome, compatible SNP rate was defined as the fraction of all scored sites showing compatible genotype between *g* and *m*, and we only retained cells with $>$0.99 compatible SNP rate.

C.1.13 Clustering analysis

CH methylation data were grouped into non-overlapping 100 kb bins across the whole genome for each cell. Due to the sparsity of the snmC-seq data, few bins had sufficient coverage ($>$100 base calls) across all cells to be retained in the analysis. We therefore imputed data at bins with coverage in 99.5% or more of cells, replacing missing values with the average methylation across all cells for that bin. This allowed us to include 76.2% of bins in the mouse genome and 63.4% of bins in the human genome in our analysis.

To cluster cells, we adapted an iterative, hierarchical and unsupervised clustering method called BackSPIN that had been previously applied to single cell transcriptome data [36]. At each iteration, the top 2,000 bins with the greatest variance across cells were selected. The SPIN algorithm was then used to arrange cells in a linear order, with similar cells located near each other [148]. Next, cells were split into two new clusters at the optimal cut point, where the average correlation within the two new clusters was highest. To retain the split, at least one of the two new subclusters must have $>$ 15% increase in the average correlation value over the average of all cells in the original cluster. This procedure was applied recursively to each new cluster, and terminated when no clusters met the splitting criterion. To avoid producing clustering with too few cells for us to confidently analyze, we prevented further splitting of clusters with 50 or fewer

cells.

Because BackSPIN can produce different results depending on the initial order of cells, we ran the algorithm with 160 random initializations of the cell order. We selected the clustering that had the highest Dunn Index. The result had 23 clusters for mouse and 40 clusters for human. Initial inspection of the cluster results using tSNE revealed that one of the human clusters, which comprised 44 cells, was highly dissimilar from the other clusters. Cells in this cluster had little detectable mCH (global median: 0.0104), significantly lower than the cluster with the next lowest mCH level (0.0201) and the median across all cells (0.0438). We surmised that this cluster may correspond to non-neuronal cells, and we therefore excluded these cells from subsequent analysis.

To conservatively define neuronal cell types based on robust and biologically interpretable differences in DNA methylation, we next merged clusters with highly similar mCH patterns. Our heuristic choices of criteria for merging clusters does impact the final number and configuration of clusters. Rather than try to accurately determine how many cell types exist in each species, our emphasis was on using consistent clustering parameters and methods in both human and mouse to allow a rigorous cross-species comparison.

To do this we defined a set of mCH marker genes. For this analysis we profiled the mCH level across all gene bodies for each cell, requiring coverage of at least 100 CH bases. We retained genes that were covered in $\geq$20% of cells in each of the clusters, and in $\geq$50% of cells in at least one cluster. We further required coverage in at least 10 cells for each cluster. We then combined reads from all cells in each cluster to estimate the mCH level for each cluster at each gene; these mCH levels were then normalized by the average over all cells at each gene. Marker genes for each pair of clusters were defined as those which were strongly hypomethylated (mCH in the bottom 2nd percentile) in one cluster and hypermethylated (mCH above the 80th percentile) for the other cluster. The top 10-20 marker genes with largest methylation difference were identified for each pair of clusters. We then tested the statistical significance of the difference in normalized methylation between the two clusters (2-sample t-test, one-sided, $p < 0.05$). If any pair of clusters

was separated by fewer than 7 significant marker genes, we merged the pair of clusters with the fewest markers and repeated the procedure (define marker gene, test significance). This process was continued until all cluster pairs had at least 7 marker genes with significantly different mCH.

For visualization purposes, we performed dimensionality reduction using t-Stochastic Neighbor Embedding (tSNE) [129], reducing all cells to a point in 2D space. TSNE requires a perplexity parameter that is analogous to how many nearest neighbors to consider in manifold learning algorithms. We examined results using a range of perplexity values (10-1000) and found largely consistent patterns for all perplexity values $>$50 (Appendix C.3G), and all tSNE visualizations shown in this study used perplexity = 150. Importantly, our tSNE results were only for visualization purposes and did not affect the clustering of neurons, although it strongly agreed with the clusters we identified using BackSPIN adapted for DNA methylation data. To illustrate how mCH levels of marker genes vary across individual cells and clusters, we computed gene body methylation as the average mCH level of annotated genic region (from TSS to TES) (Fig. 2E,F) and normalized across cells by dividing by the mean of all cells (Appendix C.6 - C.7).

C.1.14 Validation of clustering

We examined the robustness of our neuronal clusters with respect to several experimental and analytic parameters (Appendix C.3). First, we downsampled the number of reads per cell to 10%, 20% and 40% of the full dataset using samtools view -s, followed by tSNE (Appendix C.3A). To examine whether CG methylation could be used to determine cell types consistent with those estimated using mCH, we summarized CG methylation into 100kb bins followed by tSNE (Appendix C.3F). Because CG sites are more sparse than non-CG sites, we lowered the coverage cutoff to $>$ 20 base calls for this analysis.

Because backSPIN can produce different clustering outputs given different input order of cells, we compared 200 backSPIN results with independently randomized initializations against our identified neuronal clusters. We did not perform marker-gene based merging on shuffled data

as performed to obtain our original clustering, and consequently, we would expect some level of difference between our clusters and the shuffled runs. Using the adjusted Rand index [149] and adjusted mutual information [150] to quantify similarity, we found clusterings produced from shuffled inputs were highly consistent with our final clustering for mouse and human (Appendix C.3H-I). The adjusted Rand index was more variable in human, likely because we had to merge more clusters in the original backSPIN output to obtain our final human clustering.

To quantify how read downsampling affected the presence of our neuronal clusters, we downsampled the number of reads per cell. We quantified the presence of our neuronal clusters in the downsampled results using the inverse of the Davies-Bouldin index [151], mean Silhouette coefficient [152], and Calinski-Harabasz metrics [153] (all implemented in the MATLAB function, evalclusters; Appendix C.3J-L, left). These metrics reflect the separation between clusters, relative to the variability within each cluster. All three measures showed that cluster quality remains consistently high even with 20-40% of the full reads, corresponding to an average of 280,000-560,000 mapped reads per cell. Cluster quality declines upon further downsampling to 10%. We also used these metrics to quantify how well CG methylation can recapitulate our CH-defined clusters (Appendix C.3J-L, right). The quality of clusters is similar when using CG or CH methylation, and in both cases it is significantly greater compared to a shuffled control in which cells were randomly re-assigned to a different cluster. Finally, we applied density-based clustering (DBSCAN, [131]) to the data using the tSNE coordinates as input, and found generally consistent, though not identical, results compared with backSPIN (Appendix C.3M-N). For DBSCAN, we chose parameters that produced clusters most consistent with the visual separation of cells in the tSNE space (epsilon of 0.6 in mouse and 0.8 in human; minimum points of 5 for mouse and 10 for human).

Next, to examine how many cells are required to identify neuronal clusters, we ran tSNE on a random subsample of 500 or 1,000 cells (Appendix C.3B). Even with as few as 500 cells, the cell type structure is clearly present in the tSNE output and there is little mixing of different cell

types. As expected, reducing the number of cells has the greatest impact on the least numerous cell types. We also examined how reducing (10kb) or increasing (1Mbp) the bin size, and thus changing the scale of corresponding genome features, affected the clustering results (Appendix C.3C). Although tSNE results are altered at these two binning levels, the overall cell type structure is still present.

Furthermore, we examined whether mCH information from intra- or inter-genic regions is sufficient to estimate neuronal cell types. After including only reads from within gene bodies (intragenic) or which fall at least 10kb away from the nearest gene body (intergenic), we repeated the binning and tSNE procedure and found similar results (Appendix C.3D). There is therefore sufficient information in both genic and intergenic compartments for cell type classification, although we did find that the ratio of inter-cluster variance to intra-cluster variance for individual genomic bins is generally larger for genic regions (Appendix C.3E).

Finally, we examined the relationship between experimental factors (e.g. batches, random primer index) and our clusters to identify any potential experimental confounds. We used a chi-squared test for categorical variables and an ANOVA with scalar variables. Clustering was not significantly associated with experimental factors (adjusted p-value > 0.1, Appendix C.5).

C.1.15 Processing of single cell and nucleus RNA-seq datasets

Single cell RNA-seq dataset of mouse somatosensory cortex was downloaded from NCBI GEO accession GSE60361 [36]. Single cell RNA-seq dataset of mouse visual cortex was downloaded from NCBI GEO accession GSE71585 [37]. Processed data (transcripts per million table) of single nucleus RNA-seq of human cortex was downloaded from http://genome-tech.ucsd.edu/public/Lake_Science_2016/ [39]. Mouse single cell RNA-seq datasets were mapped to gencode VM10 reference followed by computing TPM (transcripts per million) for each annotated genes using RSEM 1.2.3 rsem-calculate-expression [154].

C.1.16 Cross-species comparison of single neuron clusters

For comparing a given human neuron cluster to mouse clusters, we computed cross-specific spearman correlation, for mCH level of marker genes showing homology between the two species [134]. Correlations were computed between gene mCH level of each individual human neuron and median gene mCH level of each mouse cluster (e.g. mL2/3). Marker genes were identified as described above - marker genes for each pair of clusters were defined as those which were strongly hypomethylated (mCH in the bottom 2nd percentile) in one cluster and hypermethylated (mCH above the 80th percentile) for the other cluster. The homologous mouse neuron type for a single human neuron was defined by the mouse cluster showing strongest correlation of gene mCH level with the single human neuron. The process effectively assigns each human single neuron to a most likely mouse homologous cluster. Comparison of mouse neuron clusters to human neuron clusters were performed similarly.

C.1.17 Comparison of single cell clusters defined by single cell/nucleus RNA-seq and single cell methylome

For comparing a given neuron cluster defined by DNA methylation to clusters defined by RNA-seq, we computed the Spearman correlation between marker gene mCH level (average mCH across annotated genic region) of each individual neuron and the median gene expression level (TPM) for each cluster defined by RNA-seq. Each single neuron was assigned to the cluster defined by RNA-seq showing minimum correlation coefficient since gene body mCH and transcripts abundance are generally inversely correlated.

C.1.18 In situ hybridization (ISH) and image analysis

Wild type 8wk old C57BL/6J male mice (Jackson Laboratory) were anesthetized with isoflurane and brains were removed. Mouse brains were fixed in 10% neutral buffered formalin

for 16 hours at room temperature, and were subjected to paraffin embedding at the UCSD Moores Histology & Sanford Consortium Histology Core lab. The sections were cut by at 5 μm thickness and mounted onto Superfrost Plus Slides (Thermo Fisher) and baked at 60°C for 1 hour. Double ISH was performed using RNAscope technology by Advanced Cell Diagnostics Pharma Assay Services. ISH slides were imaged with Olympus VS120 Virtual Microscopy Slide Scanning System using a 20x objective. Images in TIFF format were extracted using ImageJ BIOP VSI-Reader plugin [155]. Images were analyzed using a custom Matlab script. Pixel intensities were first centered around zero by subtracting the average pixel intensity from the entire image. Since RNAscope? assay labels individual RNA molecules, the overlap between co-stained probes was computed at the cell body level. In order to identify neuronal cell bodies, a 30 x 30 pixel sliding window (equivalent to 9.7 x 9.7 μm),similar to the size of a neuronal cell body, was used to scan the image with a step size of 10 pixels. Average pixel intensity was quantified for each sliding position and was standardized by converting to z-score. Probe specific z-score thresholds (Sulf1 - 3, Tle4 - 3, Adgra3 - 7, Pvalb - 7) were defined for the selection of sliding window positions that overlapped with a cell body and showed fluorescent intensity greater than the threshold. Connected sliding window positions were merged to create a list of regions of interests (ROIs), each corresponding to a cell body. The overlap between ROIs of the two imaging channels were counted to determine the co-expression of probed genes.

C.1.19 Identification of CG-DMR and superenhancer-like large CG-DMRs

Files containing unmethylated and methylated cytosine base calls for each cytosine position (allc files) were merged across single cells within each cluster to generate aggregate methylation data for each neuronal cluster. For each CpG site, base calls for the two cytosines located on opposite strands were combined to increase the power of DMR calling. CG-DMRs were then called using Methylpy DMRfind with false discovery rate cutoff = 0.01. Differentially methylated sites (DMSs) located within 250 bp of one another were combined into differentially

methylated regions (DMRs). DMRs containing at least two DMSs were retained for subsequent analyses.

For the identification of large CG-DMRs, CG-DMRs were first merged allowing 1kb distance between each other using bedtools merge -d 1000. Merged CG-DMRs with size greater than 5kb were considered large CG-DMRs. Large CG-DMRs were ranked by their size in Tables S7-8.

C.1.20 Comparison of single cell methylome methods

scBS-seq dataset was downloaded from NCBI GEO accession GSE56879 [126], scM&T-seq dataset was downloaded from NCBI GEO accession GSE74535 [81]. Since sc-WGBS data deposited to NCBI SRA contains non-redundant mapped reads from [127], we were not able to determine mapping rate and library complexity using our processing pipeline. scWGBS libraries were generated in-house from single mouse cortical nuclei using Illumina Truseq Methylation kit as described in [127]. Sequencing reads were first trimmed to remove sequencing adaptors using Cutadapt 1.11 [143] with the following parameters in paired-end mode: -f fastq -q 20 -m 62 -a AGATCGGAAGAGCACACGTCTGAAC -A AGATCGGAAGAGCGTCGTGTAGGGA. For singleplex samples, -m parameter was set to 40. 10 bp were further trimmed from both 5'- and 3'- ends of R1 and R2 reads to remove random primer index sequence with the following parameters: -f fastq -u 16 -u -16 -m 30. R1 and R2 reads were mapped separately as single-end reads using Bismark v0.15.0 with parameter –bowtie2. For mapping scBS-seq data, –pbat option was activated for mapping R1 reads [144]. For mapping sc-WGBS data, –pbat option was activated for mapping R2 reads. Library complexity was estimated using R1 reads with Preseq gc_extrap function with options -e 5e+09 -s 1e+07 [156].

To determine the enrichment of CpG islands (CGI) in single cell methylome data, the fraction of CGI on mouse chromosome 1 covered by a single cell methylome was compared to shuffled regions with matching sizes. The shuffling was carried out using bedtools shuffle

and was repeated five times and the average fraction of regions covered by reads was used. Bulk MethylC-seq data was downsampled to 1 million non-clonal reads for this anlsysis. For computing the amount of genomic regions covered by reads at different sequencing coverage, 1kb and 10kb bins were generated using bedtools makewindows across mouse genome. The bins were intersected with bulk MethylC-seq and single cell methylomes downsampled to 100,000 to 1 million reads.

C.1.21 Transcription factor (TF) binding motif enrichment analysis

TF binding motif enrichment analysis was performed as described in [5, 157] with the following modifications. The analysis of TF binding motif enrichment in mouse and human CG-DMRs only considered TFs with median TPM $\geq$ 10 in any clusters defined by single cell/nucleus RNA-seq of mouse visual cortex and human cortex, respectively [37, 39]. To summarize enriched or depleted TF binding motifs to TF classes, classification of TFs was downloaded from http://tfclass.bioinf.med.uni-goettingen.de/tfclass [158]. The folds of enrichment or depletion for TF classes were defined as the strongest enrichment or depletion shown by TF class members.

C.1.22 Prediction of putative enhancers

The enhancers of three major brain cell types (excitatory neurons, PV neurons and VIP neurons) were predicted using Regulatory Element Prediction based on Tissue-specific Local Epigenetic marks (REPTILE) [135]. REPTILE integrates DNA methylation and chromatin accessibility data to delineate the location of enhancers. REPTILE formulates enhancer prediction as a supervised learning task - it learns the chromatin signatures of enhancers (i.e. enhancer model) using known enhancers and then makes predictions across the whole genome in various cell types and tissues. A unique feature of RETPILE is that it is able to incorporate the data of cells and tissues besides the target sample (as outgroup) and utilize the variation of epigenomic

data to improve prediction accuracy.

To generate the putative enhancers of the three brain cell types, we first downloaded the bulk WGBS and ATAC-seq data of excitatory neurons (EXC), PV neurons (PV) and VIP neurons (VIP) from Gene Expression Omnibus (GEO). The accessions of ATAC-seq data are: EXC (GSM1541964, GSM1541965), PV (GSM1541966, GSM1541967) and VIP (GSM1541968, GSM1541969). The accessions of WGBS are EXC (GSM1541958, GSM1541959), PV (GSM1541960, GSM1541961) and VIP (GSM1541962, GSM1541963). In order to train a mouse enhancer model, we also obtained the EP300 ChIP-seq data (GSM723018) and its corresponding input (GSM723020) in mouse embryonic stem cells (mESCs) as well as the bulk ATAC-seq data (GSM2156965) and bulk WGBS (GSM1162043 and GSM1162044) of mESCs.

Next, The WGBS and ChIP-seq data were processed as previously described [135]. ATAC-seq data were processed in the same way as [5]. Data of replicates were combined. EP300 binding sites were identified using MACS2 [159] similar to [135]. DMRs were called across the methylomes of mESCs and three brain cell types as previously stated [135].

Then, we trained a mouse enhancer model in mESCs. The construction of training dataset as described in [135] - the EP300 binding sites were treated as positive instances, whereas promoters and randomly chosen genomic bins were used as negative instances. During the training process, the data of three brain cell types were used as outgroup. After training, we obtained a mouse enhancer model, which is able to distinguish enhancers from genomic background based on the mCG and open chromatin signatures.

Lastly, we applied this model to generate enhancer predictions for three brain cell types. During this process, when REPTILE made enhancer predictions for one brain cell type, mESCs and the other two brain cell types were used as outgroup.

C.1.23 Prediction of excitatory neuron super-enhancers

Excitatory neuron super-enhancers were identified with ROSE ([137]) using the list of H3K27ac peaks identified from cortical excitatory neurons with parameters -s 12500 -t 2500. Excitatory neuron H3K27ac ChIP-seq data and peaks were reported in [5].

C.1.24 Comparative analysis of regulatory elements

CG-DMRs were categorized based on the conservation of sequences and methylation states between human and mouse. First, UCSC liftOver [140] was used to project CG-DMRs between species based on sequence conservation (minimum ratio of remapped bases = 0.1). CG-DMRs that could be mapped to the other species and mapped back were referred to as mappable CG-DMRs. All other CG-DMRs were called unmapped DMRs. Next, we further divided mappable CG-DMRs into two categories: CG-DMRs located within 1kb of a CG-DMR in the other species (shared CG-DMRs), and CG-DMRs located further than 1 kb away from any DMR in the other species after liftover (specific CG-DMRs).

We then used a hypergeometric test to examine whether mappable CG-DMRs from one species preferentially overlap with CG-DMRs in the homologous cell type from the other species. Specifically, to calculate the expected number of shared DMRs between human cluster i and mouse cluster j, we mapped (via liftover) human cluster i DMRs to the mouse genome and found how many were shared with any mouse DMR (i.e. overlap within 1kb of merged mouse DMRs, N_{ij}). Then this number was divided by the total number of merged human DMRs (N_h) and multiplied by the number of DMRs in human cluster i (N_{hi}), to get the expected number of shared DMRs: $E_{ij} = \frac{N_{ij}}{N_h} N_{hi}$. This was compared with the observed number of shared DMRs between human cluster i and mouse cluster j, N_{ij}.

To quantify the regulatory conservation of CG-DMRs between human and mouse, we computed the correlation of methylation levels at mappable DMRs for each homologous cluster

pair. The higher cross-species correlation of inhibitory clusters suggests that inhibitory neurons have greater regulatory conservation between the two species (Fig. 4E, Appendix C.17C, $p < 0.001$, Wilcoxon rank-sum test). This finding was further corroborated by examining cross-species enrichment of shared CG-DMRs represented by the fold-change between observation and expectation, which again showed stronger overlap between the two species in inhibitory than excitatory neuron clusters (Appendix C.17D-E).

To measure sequence conservation of CG-DMRs, we computed 100-way PhastCons score for human CG-DMRs and 60-way PhastCons score for mouse CG-DMRs by taking the average PhastCons score across all bases in each DMR. Missing values were skipped rather than treated as zero. We observed higher PhastCons scores at CG-DMRs in inhibitory neuron clusters than in excitatory (Fig. 4E, $p < 0.001$, Wilcoxon rank-sum test), suggesting greater sequence conservation of inhibitory neuron regulatory elements across species. We then performed the same analysis at putative enhancers identified by REPTILE from purified neuronal populations [135], and also found greater sequence conservation in inhibitory than excitatory neuron putative enhancers (Appendix C.17F).

We calculated average PhastCons score across the region surrounding (+/- 10kb) of excitatory or inhibitory neuron CG-DMRs with 100 bp resolution. The conservation of inhibitory neuron CG-DMRs is greater than in excitatory only within 500bp around the CG-DMR (Appendix C.17G). Then we investigated whether genes preferentially expressed in inhibitory neurons are more conserved at their gene body. We used nuclear RNA-seq data [5] to find genes with at least 2 fold over-expression in one cell type against the other two cell types. Excitatory neuron specific genes showed greater sequence conservation than PV and VIP specific genes at their TSS and similar conservation in flanking regions (Appendix C.17H). This result indicates that the higher conservation of inhibitory cells may be restricted to the regulatory elements.

We further divided mappable CG-DMRs into two categories: those proximal to a TSS (within 25kb) and those distal to a TSS, computing the PhastCons score for each. Results showed

greater conservation of inhibitory neuron CG-DMRs, with the difference being more pronounced for distal CG-DMRs (Appendix C.17I).

Finally, we examined whether excitatory and inhibitory neuron CG-DMRs associated with the same gene (within 25kb of TSS) show different conservation. For each gene, we compared the PhastCons score between DMRs in excitatory cells and inhibitory cells that associated with the genes, and again we observed a significant higher conservation in inhibitory DMRs (Appendix C.14J, $p < 1e-6$, Wilcoxon signed-rank test).

C.2 Supplementary texts

C.2.1 snmC-seq shows reliable sample multiplexing and high reads mapping rate

Our snmC-seq protocol starts from separating single nuclei using fluorescence-activated cell sorting (FACS) and dispense into wells of 384- well PCR plates followed by proteolytic digestion and bisulfite conversion. snmC-seq is compatible with both fresh and frozen tissues. We incorporated 5'- sequencing adaptors through indexed random primer-initiated DNA synthesis. (Fig. 1A) After pooling four indexed random priming reactions, 3'- sequencing adaptors were incorporated using AdaptaseTM technology. The majority of multiplexed pools were constructed by combining two mouse and two human nuclei. Mapping of sequencing reads to both mouse and human reference genomes showed negligible cross-species mapping (Appendix C.2A), confirming the fidelity of the multiplexing strategy.

We have compared snmC-seq with published methods for single cell methylome including scBS-seq [126], scWGBS [127], scM&T-seq [81]. We specifically examined fraction of reads retained after adaptor trimming, unique mapping rate and library complexity (Appendix C.2B-D). We generated scWGBS libraries from single mouse cortical nuclei using Illumina Truseq

Methylation kit as described in [127]. snmC-seq shows significantly greater mapping rate (median = 52.7%) compared to scM&T-seq (median = 19.8%, [81]) and scBS-seq (median = 22.5%, Appendix C.2C [126]). The mapping rate of snmC-seq is comparable to scWGBS libraries (median = 55.4%). However, snmC-seq library contains approximately four times more unique molecules than scWGBS libraries (Appendix C.2D).

Reads coverage pattern was compared between downsampled traditional bulk MethylC-seq data, snmC-seq, scBS-seq [126] and sc-WGBS [127]. It was previously shown that the coverage of CpG islands (CGI) is enriched in single cell methylome [126, 127]. CGI enrichment was determined relative to shuffled genomic regions with matching sizes (Appendix C.2E). snmC-seq showed similar enrichment of CGI (mean fold change = 1.65x) as sc-WGBS (mean fold change = 1.54x), while scBS-seq showed moderately higher CGI enrichment (mean fold change = 2.02x). Traditional MethylC-seq showed depletion of CGI with a mean fold change of 0.52x.

To quantify the evenness of single cell methylome coverage, the fraction of non-overlapping 1kb and 10kb genomic bins covered by sequencing reads were plotted as a function of sequencing depth for each method (Appendix C.2F). With a given number of sequencing reads, traditional MethylC-seq data always covers most genomic bins, suggesting less coverage bias than single methylome methods. Single cell methylome methods show moderate difference between their coverage evenness, with snmC-seq showing intermediate evenness between sc-WGBS and scBS-seq measured with 1kb bin coverage, and near identical evenness with sc-WGBS measured with 10kb bin coverage.

C.2.2 hPv-2 is a potentially human specific PV+ inhibitory neuron population

hPv-2 represented a potential human-specific inhibitory population. The strong hypomethylation of GAD1 and LHX6 genes in hPv2 suggests these are inhibitory neurons derived from the medial ganglionic eminence (MGE) (Appendix C.12B); however, hPv-2 was the only

inhibitory neuron cluster in either mouse or human showing hypermethylation of GAD2. Notably, hPv-2 cells have low mCH at CCK and high methylation at GRIK3, similar to caudal ganglionic eminence (CGE) derived interneurons, such as VIP cells, and distinct from MGE-derived inhibitory cells.

Unique large CG-DMRs were also found in hPv-2 (Appendix C.16E,J). Gene bodies of NACC2, UNC5B, FAM20C and FAM222A were demethylated in hPv-2 but not in any other human or mouse inhibitory neuron clusters (Appendix C.16E, J). Thus, these observations suggest hPv-2 is a unique human PV neuron population defined by both marker gene mCH and super-enhancer mCG signatures.

C.2.3 Inhibitory neurons show layer-specific DNA methylation signatures

We found that global mCH level differed among inhibitory neurons within a clusters but located in different cortical layers (Appendix C.14A). For example, PV+ interneurons located in superficial layers had significantly less global mCH than middle and deep layer PV+ neurons ($p < 1x10^{-5}$, Wilcoxon rank sum test). Significant global mCH layer differences were also found between superficial and middle layer SST+ neurons ($p < 1x10^{-3}$, Wilcoxon rank sum test). Moreover, we identified 406 genes with layer specific mCH in PV+ neurons (Appendix C.14B, one way ANOVA q-value < 0.01; Table S4). The vast majority (358) of these genes were hypomethylated in superficial layer PV+ neurons. In addition, MGE-derived inhibitory populations, including PV+ and SST+ but not CGE-derived VIP+ neurons, shared a significant fraction of genes showing hypomethylation in superficial layers (hypergeometric p-value=$8.7x10^{-59}$, Appendix C.14E), suggesting that layer-specific gene regulation in mature inhibitory neurons may be defined by their progenitor zones. Genes with low mCH in superficial layer PV+ neurons are enriched in functional annotations including neurogenesis, axon guidance functions and synapse part (Appendix C.14F-H), suggesting layer-specific epigenetic regulation of synaptic functions in inhibitory neurons.

We identified human PV+ and SST+ neurons that putatively located in different layers by comparing to mouse superficial or deep layer neurons (Appendix C.14I). Human PV+ and SST+ neurons putatively located in different layers were separated by tSNE (Appendix C.14J and K). Superficial and deep layer human SST+ neurons were also separated by clustering, with hSst-2 correlated with superficial layer mSst-1 whereas hSst-1 and hSst-3 correlated with deep layer mSst-1 (Appendix C.14I). Superficial layer human PV+ and SST+ neurons had less global mCH compared with neurons of the same type located in deep layers (Appendix C.14M). A group of genes, including *Cux2*, *Nlgn1*, *Grin2a* and *Shank2*, showed similar layer specific mCH patterns between mouse and human (Appendix C.14N-O).

C.2.4 Large CG-DMR is a reliable marker for superenhancer

We tested the specificity of superenhancer prediction with large DMR using CG-DMRs found in mL2/3. Putative excitatory neuron superenhancers were predicted using enhancer histone mark H3K27ac profile of purified Camk2a+ excitatory neurons with software ROSE [5, 137]. mL2/3 has a high-coverage aggregated methylome from 690 single neurons, which allowed sensitive CG-DMR calling for this cluster. We first merged mL2/3 CG-DMRs located within 1kb from one another and then ranked merged CG-DMRs by their size. We found that the enrichment of H3K27ac over merged CG-DMRs increases along with the size of CG-DMRs (Appendix C.16A), suggesting large CG-DMRs are associated with strong regulatory activity. A greater portion of large DMRs (e.g. $>$ 15kb) were overlapped with putative superenhancers, compared to DMRs with smaller sizes (Appendix C.16B). For example, 90.3% of merged DMRs larger than 15kb, whereas only 24.8% of merged DMRs with size greater than 2kb were overlapped with putative superenhancers.

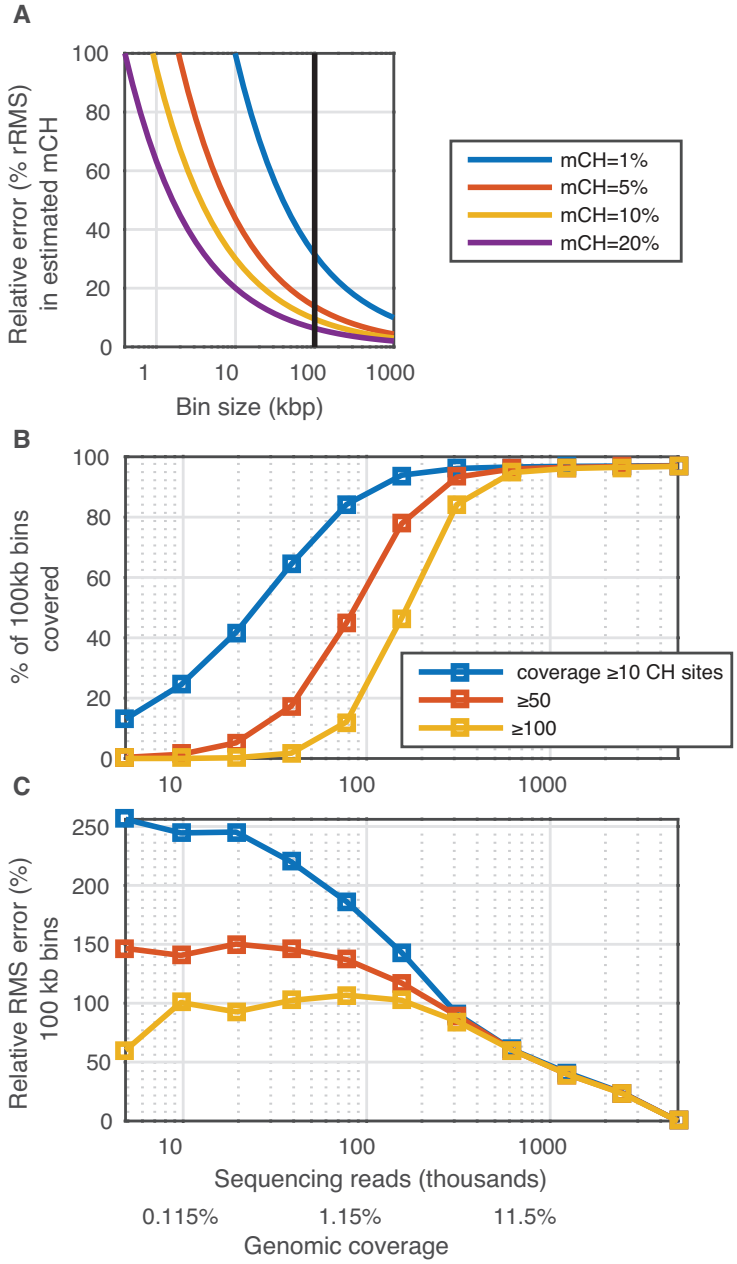

Figure C.1: mCH can be accurately estimated within 100kb bins using sparse snmC-seq data. (**A**) Theoretical model estimates the relative RMS error in mCH in genomic bins e = sqrt((1-p)/(prcb)), where $p \approx 0-0.2$ is the true methylation level, $r \approx 0.05$ is the genomic coverage, $c \approx 0.2$ is the fraction of CH positions in the genome, and b is the genomic feature size. (**B**) Downsampling a deeply sequenced single cell methylome shows that $> 95\%$ of the genome can be covered with ≥ 100 CH basecalls per 100 kb bin, assuming $\sim$ 500,000 reads or $\sim$5% genomic coverage. (**C**) The rRMS error is estimated by comparing the mCH estimated using the full coverage data with downsampled data.

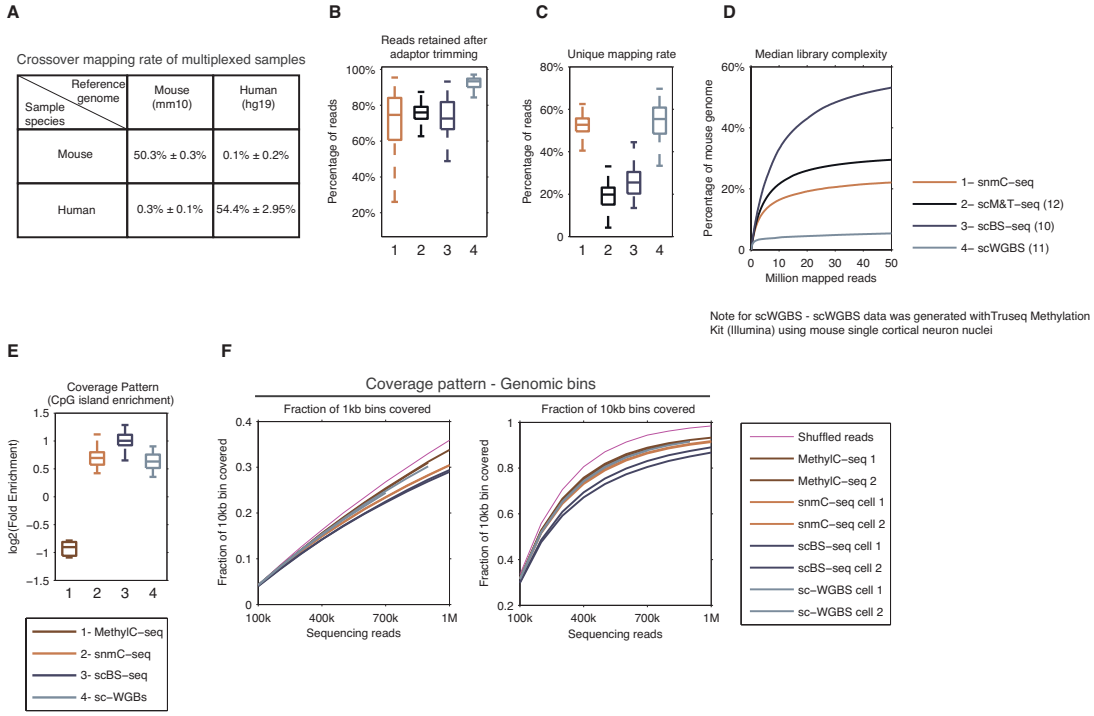

Figure C.2: snmC-seq is compatible with multiplexing and demonstrates efficient read mappability. (**A**) Mapping of 100 randomly selected multiplexed snmC-seq samples to both human and mouse reference genomes showed no species crossover between pooled indexed random priming reactions. (**B**) Percentage of sequencing reads retained after trimming of generic Illumina adaptors, random primer index and low complexity tail introduced by AdaptaseTM for snmC-seq. (**C**) Percentage of trimmed sequencing reads that were uniquely mapped. (**D**) Complexity of single cell methylome libraries estimated using R1 reads. (**E**) Enrichment of CpG islands in DNA methylome generated by traditional MethylC-seq, snmC-seq, scBS-seq and sc-WGBS. (**F**) Fraction of 1kb and 10kb non-overlapping bins covered by single cell methylome data as a function of the number of sequencing reads.

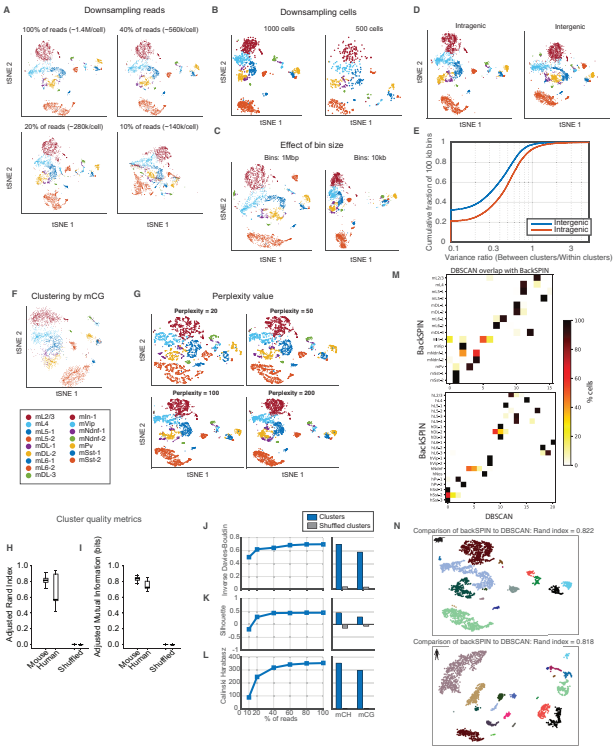

Figure C.3: Single nuclei are consistently clustered by cell type using multiple methylome features across a wide range of genomic length scales. (**A**) Cell types can be clearly separated in tSNE space despite downsampling reads to 20% of the full mouse dataset (∼280,000 uniquely mapped reads per cell); the quality of clustering begins to break down at 10% downsampling (140,000 reads). tSNE analysis was performed using mCH levels in 100kb bins (minimum coverage, 100 base calls), and each cell was colored according to the cluster identity assigned in our analysis of the full dataset. (**B**) Major cell types are well separated by tSNE analysis using as few as 500 or 1,000 cells; increasing the number of cells increases the representation of minority cell types. (**C**) Cell type clusters can be identified by tSNE analysis using mCH in bins as small as 10kb or as large as 1Mb. (**D**) Comparison of tSNE representation of mouse clusters based on mCH in intragenic regions (including all bases between each TSS and TES) vs. intergenic regions (¿10 kb away from any gene body). (**E**) The cumulative distribution over all 100 kb bins of the ratio of between-cluster variance (i.e. the variance of the mean mCH for each cluster) vs. the within-cluster variance (i.e. the variance of all cells, after subtracting the cluster mean) for inter- and intra-genic reads. (**F**) Cell type clusters can be identified by tSNE analysis using mCG in 100 kb bins. (**G**) tSNE representation of single mouse neuron methylome with different perplexity values shows consistent patterns. (H-L) The quality of backSPIN clustering is shown for mouse and human using the adjusted Rand index (**H**), adjusted mutual information (**I**), inverse of the Davies-Bouldin index (**J**), mean silhouette index (**K**) and Calinski-Harabasz index (**L**). For indices shown in (**H**-**K**), a value close to 1 indicates that clusters are well separated relative to the variability within each cluster, while a value close to 0 indicates poor cluster separation. Box plots show the distribution of each index over 200 clustering runs with random initialization, and they are compared with the results for randomly shuffled cluster assignments. (**M**-**N**) Comparison of BackSPIN clustering results to clusters generated from tSNE and DBSCAN for mouse and human.

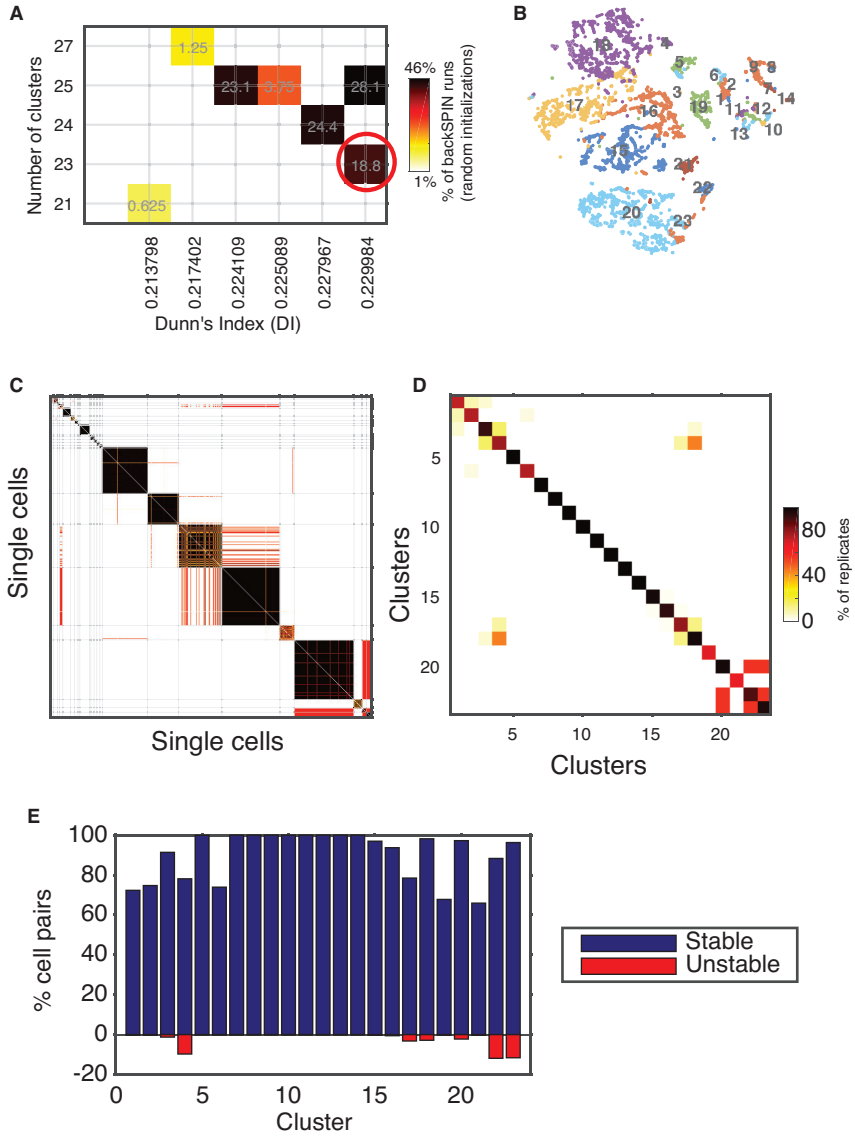

Figure C.4: Cluster robustness. (**A**) 160 independent clusterings were generated with backSPIN using random initialization. Each backSPIN run converged to one of 7 different clusterings; we selected the clustering with the highest Dunn's Index as a reference clustering (red circle). (**B**) tSNE plot showing the reference clustering. (**C**) For each pair of cells, the color shows the fraction of backSPIN runs in which those two cells were co-clustered. (**D**) Average co-clustering for all cell pairs in two different clusters. (**E**) For every cell pair in each cluster, we plot the % stability, i.e. the % of runs in which the cells are co-clustered. We also plot the % unstable, i.e. the % of cell pairs that are not in the same cluster which are co-clustered.

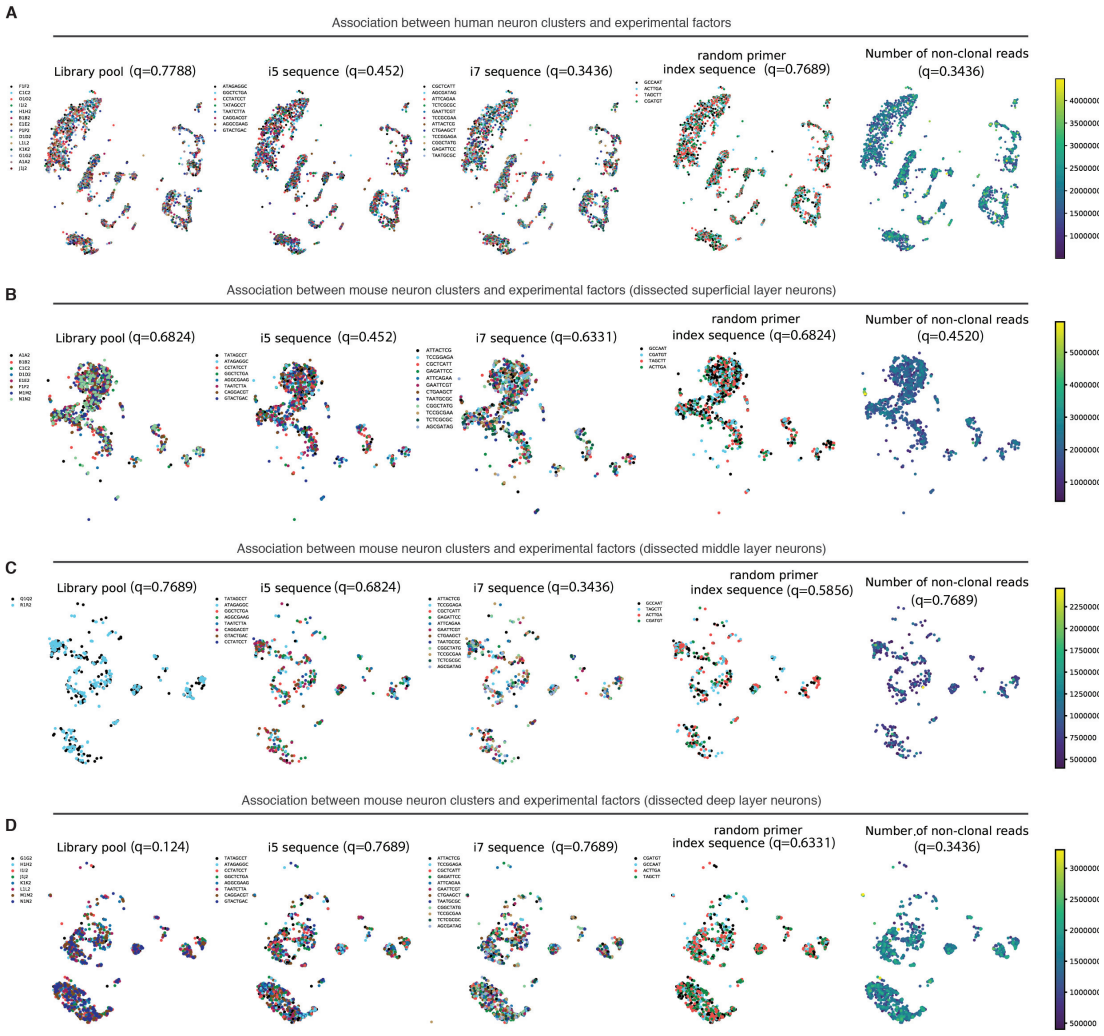

Figure C.5: Absence of strong association between neuron clusters and experimental factors. Statistical comparison of clustering to sequencing experimental factors for human neurons (**A**), dissected mouse superficial layer (**B**), middle layer (**C**) and deep layer (D) neurons.

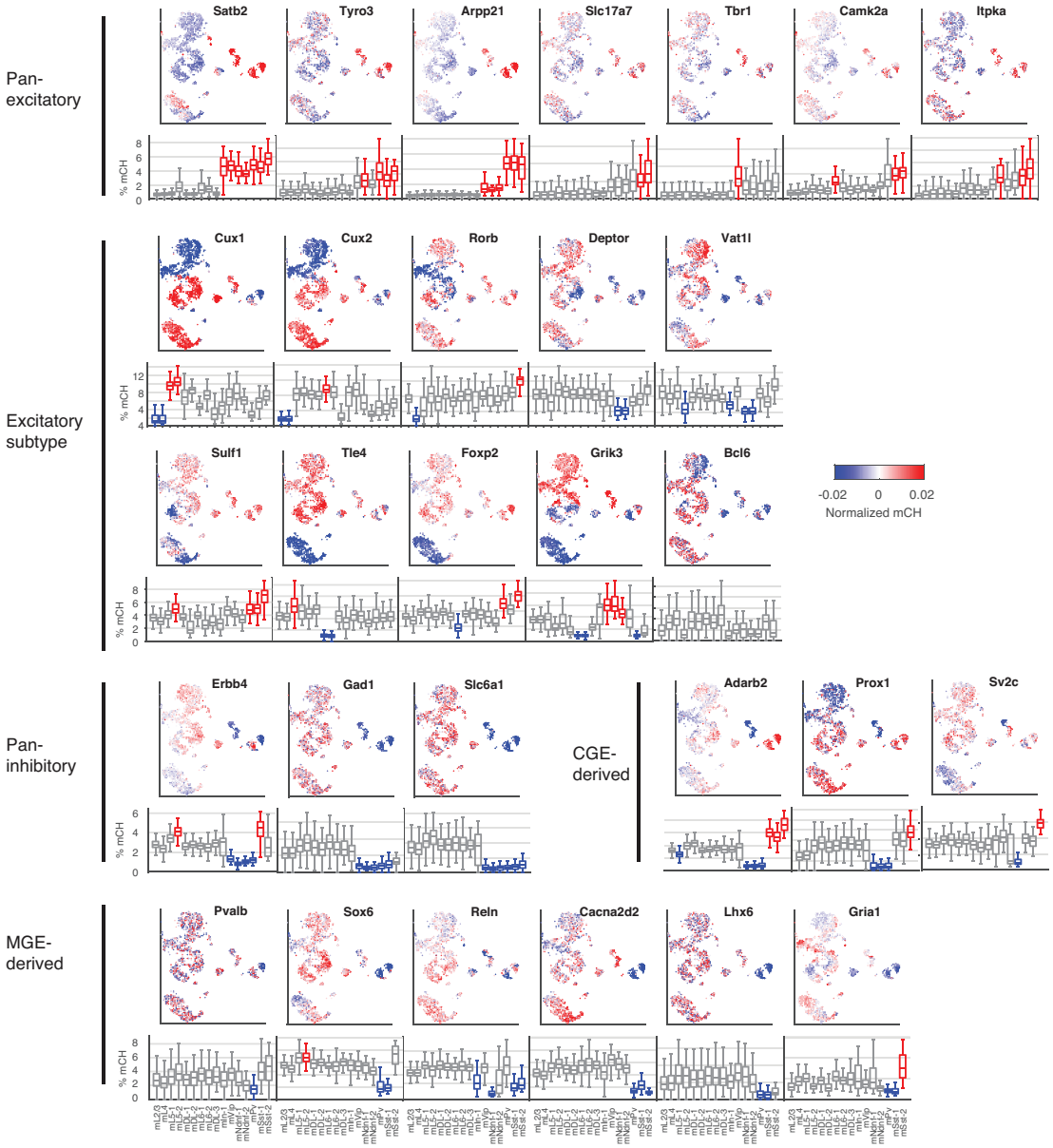

Figure C.6: Mouse marker genes. For each gene, single cells are shown in tSNE representation colored according to the normalized mCH level. Box plots below each tSNE show the distribution of absolute (not normalized) mCH level across all cells within each cluster. For each gene's box plot, we highlight clusters that are significantly hypermethylated (red) or hypomethylated (blue). Hypermethylated (hypomethylated) clusters are defined to be clusters for which at least 75% of cells have higher (lower) methylation than the top (bottom) 25% of cells in all other clusters.

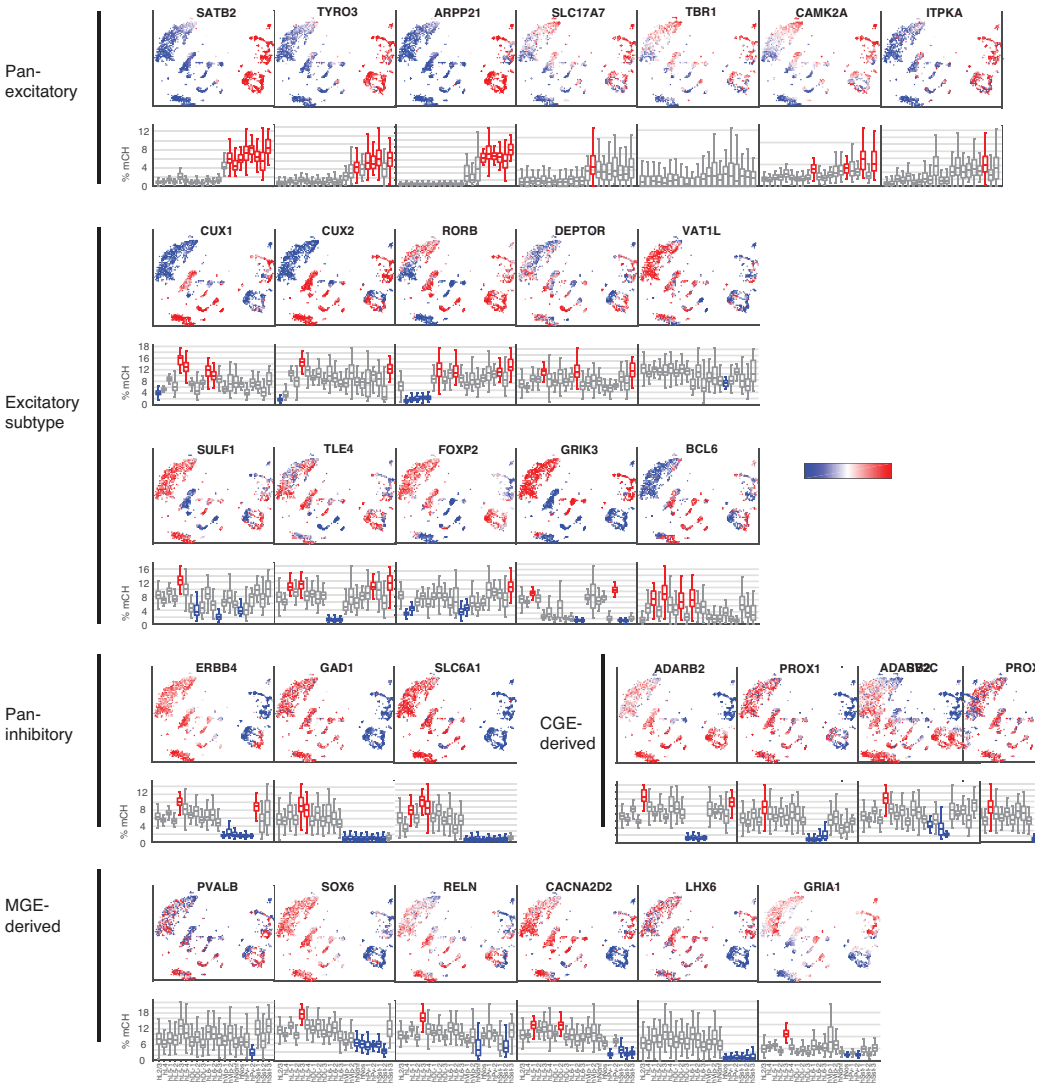

Figure C.7: Human marker genes. For each gene, single cells are shown in tSNE representation colored according to the normalized mCH level. Box plots below each tSNE show the distribution of absolute (not normalized) mCH level across all cells within each cluster. For each gene's box plot, we highlight clusters that are significantly hypermethylated (red) or hypomethylated (blue). Hypermethylated (hypomethylated) clusters are defined to be clusters for which at least 75% of cells have higher (lower) methylation than the top (bottom) 25% of cells in all other clusters.

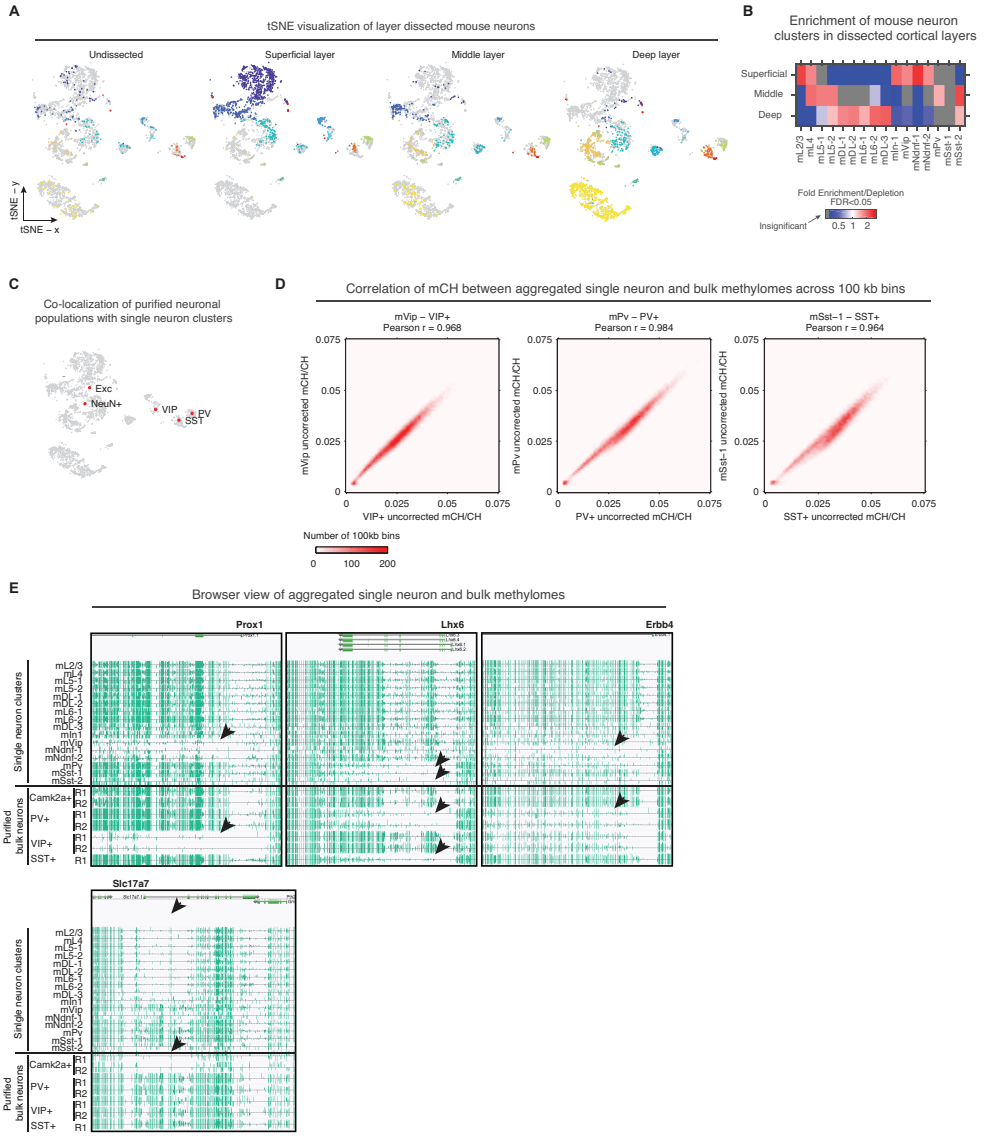

Figure C.8: Single neuron clusters are correlated with layer dissection and bulk methylome generated from purified neuron populations. (**A**) Single neurons isolated from undissected and dissected superficial, middle and deep layer frontal cortex tissue are separately visualized using tSNE. (**B**) Enrichment/depletion of cells from mouse dissected superficial, middle and deep cortical layers in neuron clusters. (**C**) Immunologically (NeuN+) and genetically labeled (Exc - Camk2a+, PV - PV+, VIP - VIP+, SST - SST+) neuron populations are co-clustered and visualized together with mouse single neurons using tSNE. (**D**) Consistent mCH profiles between bulk methylome and aggregated single neuron methylomes for nonoverlapping 100 kb bins across the mouse genome. Bulk methylome generated from purified mouse neuronal populations (VIP+, PV+ and SST+) were compared with aggregated single neurons methylomes of corresponding clusters (mVip, mPv and mSst-1). (**E**) Browser tracks showing concordance of snmC-seq data pooled from neuronal cell type clusters (top tracks) with bulk DNA methylation profiling of purified neuronal cell types. Arrows indicate corresponding single cell clusters and bulk cell types with low methylation levels at these cell-type specific loci.

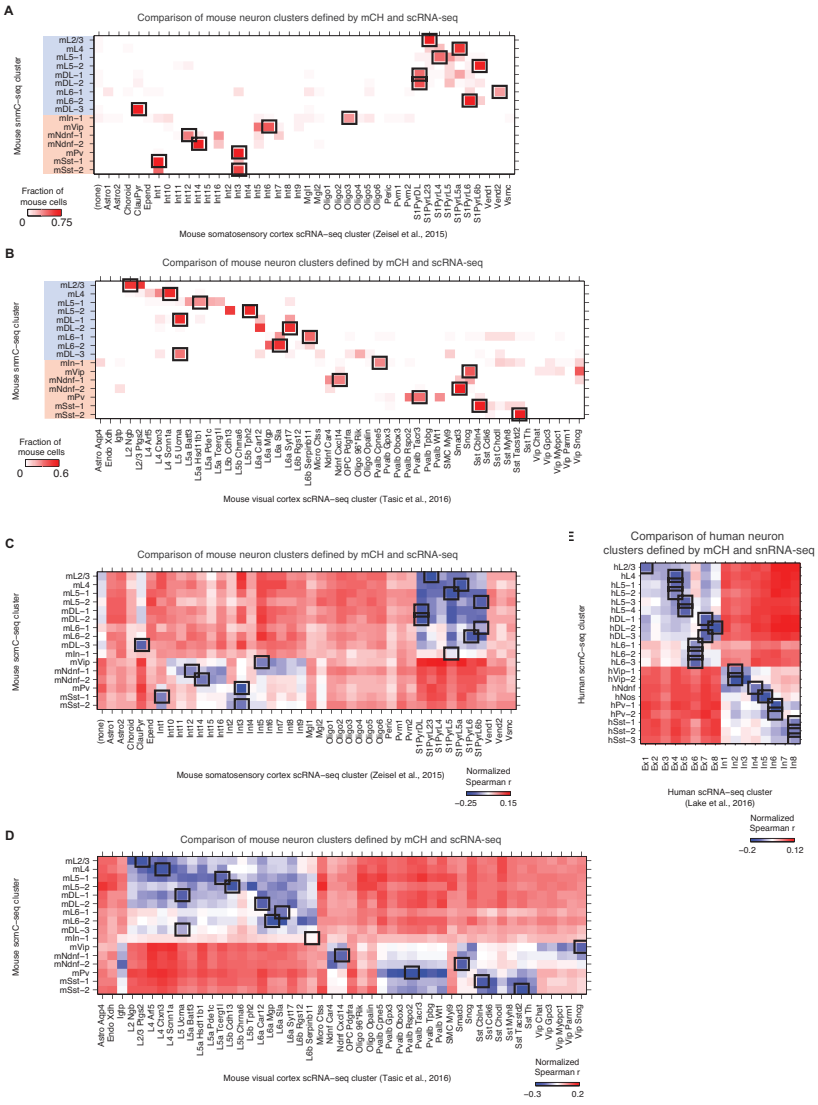

Figure C.9: Correlation between single neuron clusters defined by snmC-seq and single cell/nucleus RNA-seq. Comparison of mouse neuron clusters to mouse somatosensory cortex single cell clusters defined by scRNA-seq (**A**, [36]) and mouse visual cortex single cell clusters defined by scRNA-seq (**B**, [37]). For (A) and (B), color represents the fraction of cells in each snmC-seq cluster having the best match (strongest inverse correlation) for an RNA-seq cluster. The best RNA-seq cluster match for each of snmC-seq clusters was highlighted with a black rectangle. (**C**-**E**) Normalized Spearman correlation coefficients between gene body mCH level in snmC-seq clusters and median transcript abundance (TPM) of RNA-seq clusters. Mouse snmC-seq clusters were compared to mouse somatosensory cortex single cell clusters defined by scRNA-seq (C, [36]) and mouse visual cortex single cell clusters defined by scRNA-seq (D, [37]). (E) Human snmC-seq clusters were compared with human cortical neuron clusters defined by snRNA-seq [39]. Spearman correlation coefficients were normalized by subtracting the mean value of each row in the matrix.

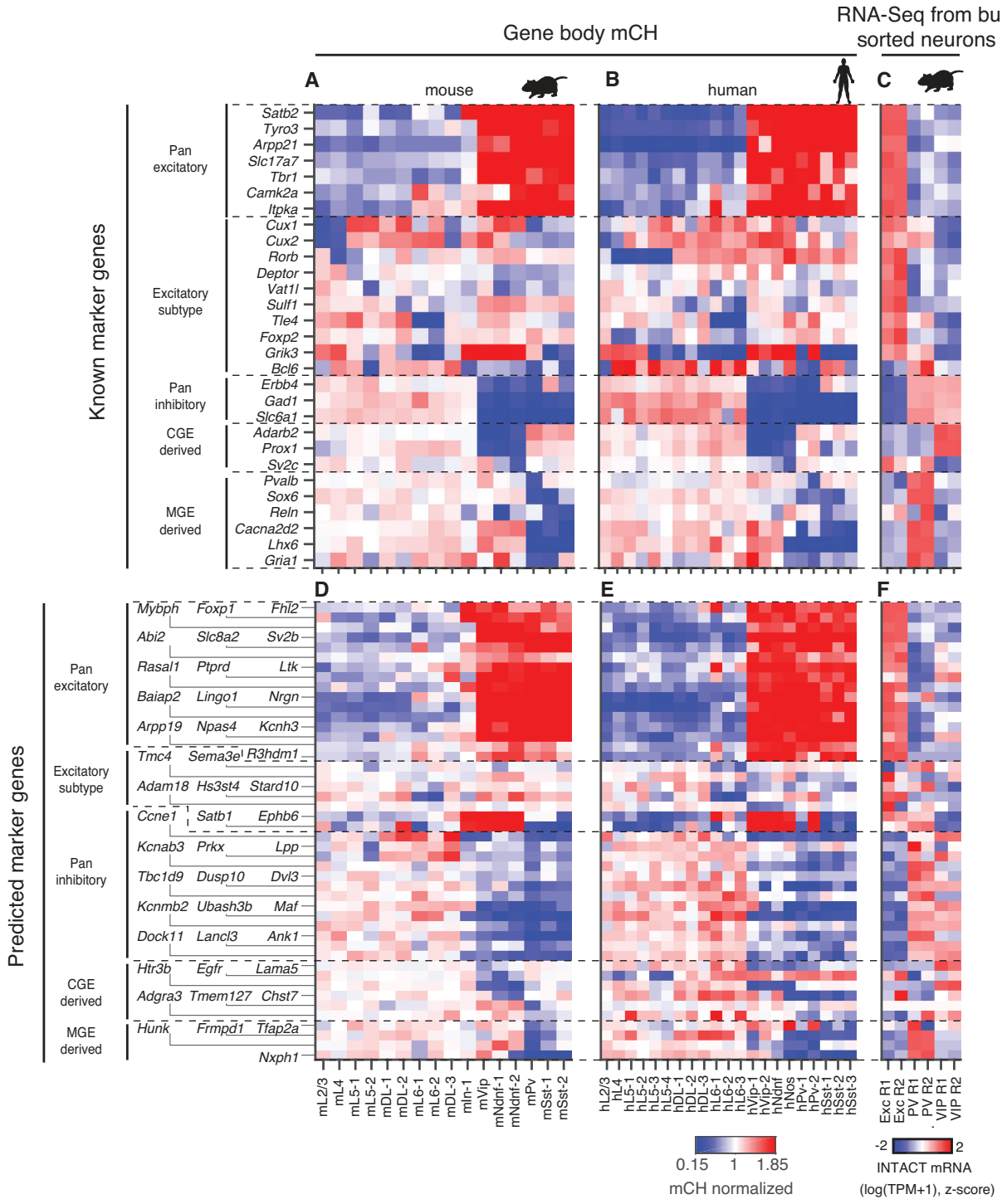

Figure C.10: Prediction of neuron type marker genes with single cell methylomes. (**A**-**B**) mCH level of known markers for mouse (A) and human (B) neuron clusters. (**C**) Gene expression of known markers for Camk2a+ (Exc), PV+ and VIP+ neuron populations. (**D**-**E**) mCH level of newly predicted markers for mouse (D) and human (E) neuron clusters. (**F**) Gene expression of newly predicted markers for Camk2a+, PV+ and VIP+ neuron populations.

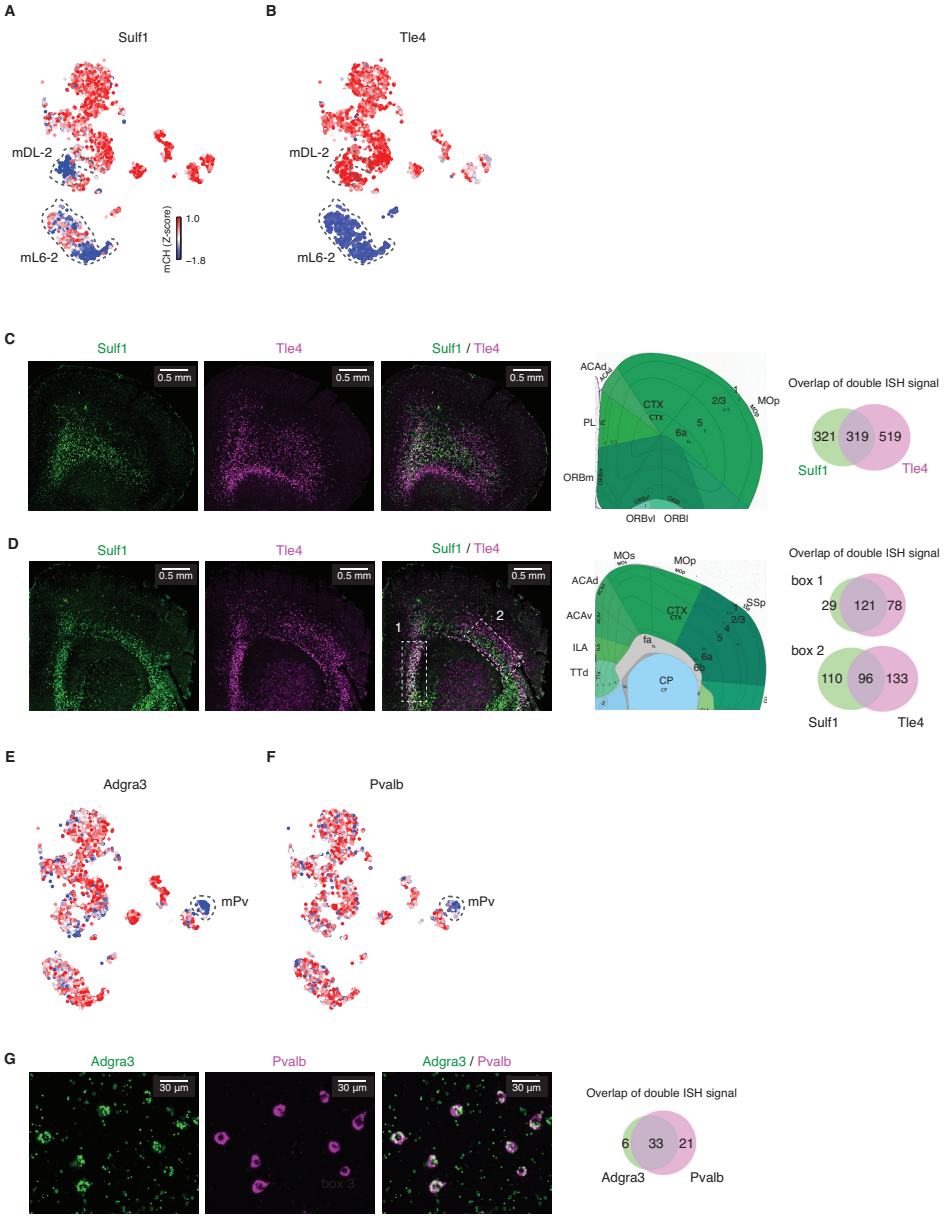

Figure C.11: Double ISH experiments validate novel markers predicted by mCH. (**A**-**B**) Relative mCH level (mCH Z-score) of Sulf1 and Tle4. The z-score is defined as the mCH value minus its mean over all cells, divided by the standard deviation across cells. (**C**-**D**) Double in situ RNA hybridization results using probes for Sulf1 and Tle4 in mouse FC. (C) and (D) show two coronal sections both in mouse FC with (C) located more rostral than (D). (**E**-**F**) Relative mCH level (mCH Z-score) of Adgra3 and Pvalb. (**G**) Double in situ RNA hybridization results using probes for Adgra3 and Pvalb.

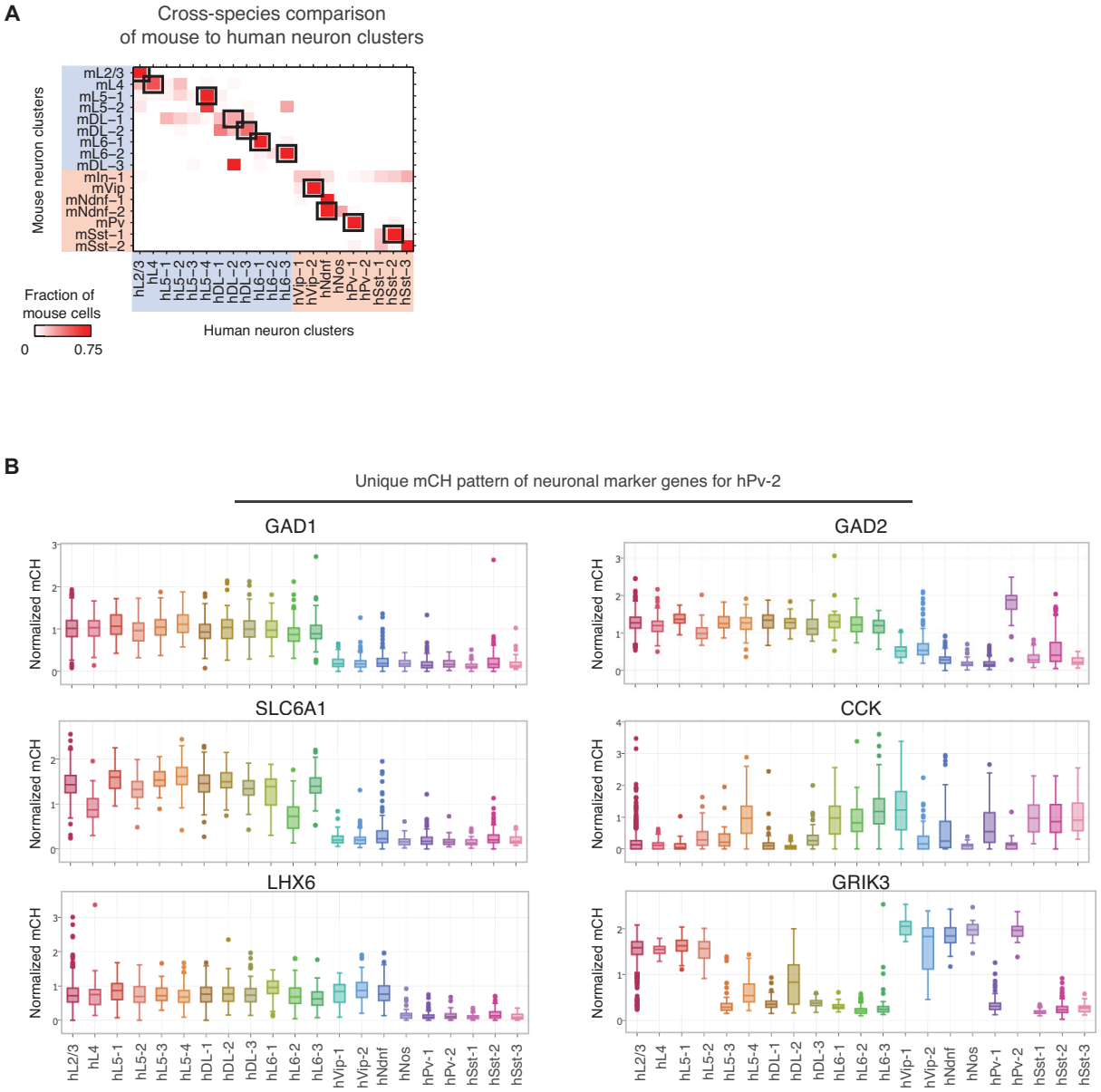

Figure C.12: Expanded neuronal diversity in human FC. (**A**) Cross-species cluster similarity computed by comparing mouse to human clusters. Color indicates the fraction of neurons in mouse cluster showing strongest correlation (Spearman correlation at homologous gene bodies) with each human cluster. Human and mouse cluster pairs that are mutual best matches are highlighted with black rectangles. (**B**) hPv-2 shows unique mCH pattern of neuronal marker genes. Boxplots show the distribution of gene body mCH of individual single human neurons, normalized by dividing gene body mCH by global mCH for each cell.

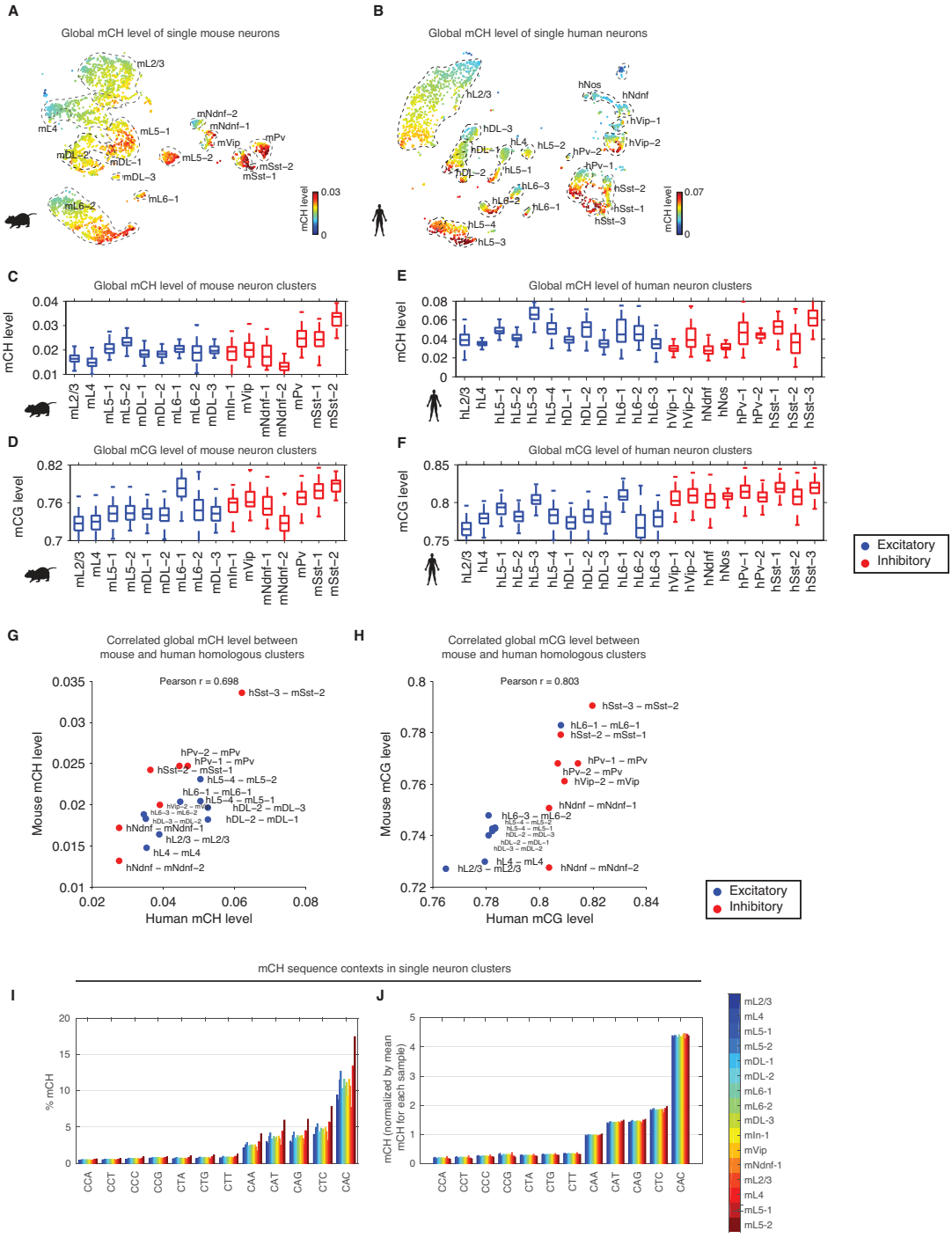

Figure C.13: Global mC levels are conserved between mouse and human neuron types. (**A** and **B**) Global mCH level for single mouse (A) and human (B) neurons. (**C**-**D**) Genome-wide mCH (C) and mCG (D) levels for mouse neuron clusters. (**E**-**F**) Genome-wide mCH (E) and mCG (F) levels for human neuron clusters. (**G**-**H**) Cross-species comparison of genome-wide mCH and mCG level between homologous clusters. (**I**) Percentage of mCH basecalls located in each trinucleotide context for mouse neuron clusters. (**J**) Normalized mCH level of each trinucleotide context for mouse neuron clusters.

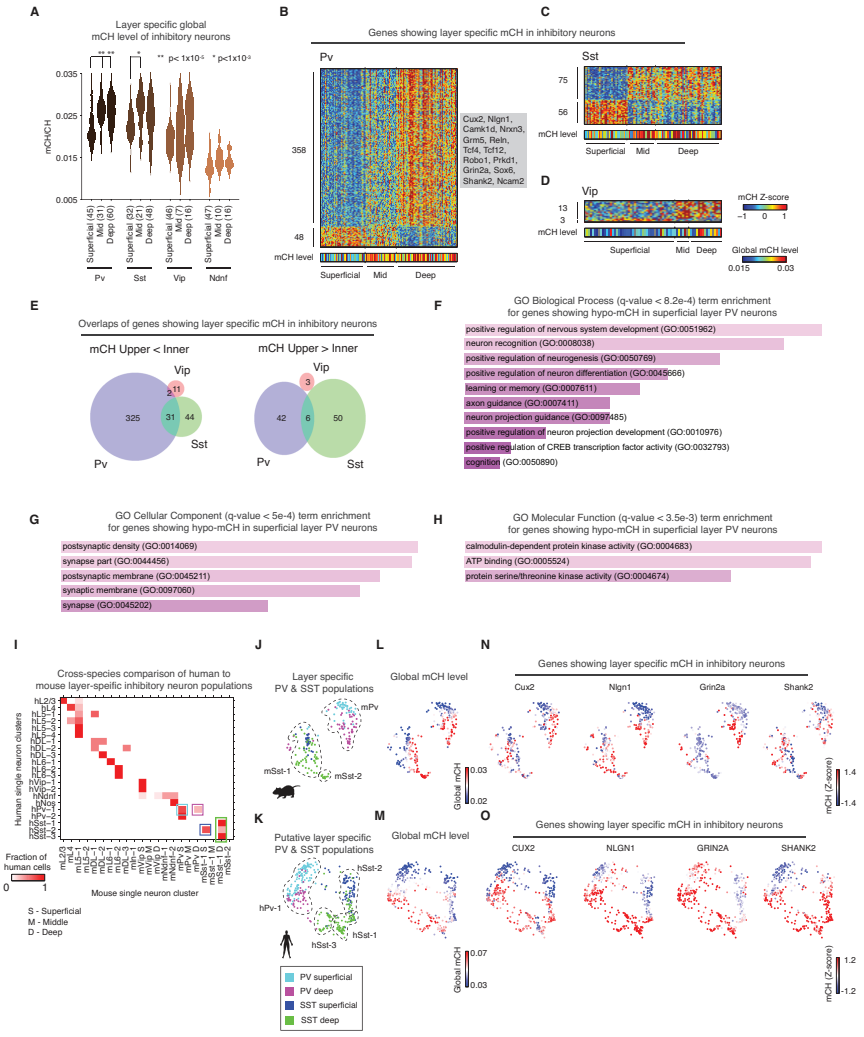

Figure C.14: Inhibitory neurons possess global and gene level cortical-layer-specific mCH signatures. (**A**) Global mCH level shows layer-specific differences in PV and SST neurons. (**B**-**D**) A subset of genes show layer-specific gene body mCH in inhibitory neurons. (**E**) Overlap of genes showing layer-specific mCH in PV and SST neurons. (**F**-**H**) Gene ontology term enrichment for genes showing hypo-mCH in PV neurons located in superficial layer. (**I**) Cross species similarity computed by comparing human to mouse clusters, with mouse inhibitory clusters divided into sub-clusters containing neurons located in dissected superficial, middle and deep cortical layers. Cyan rectangle indicates human neuron showing strongest correlation to mouse superficial layer PV neurons, magenta rectangle indicates correlation to mouse deep layer PV neurons, blue rectangle indicates correlation to mouse superficial SST neurons, and green rectangle indicates correlation to mouse deep SST neurons. (**J**) tSNE visualization of mouse PV and SST neurons located in superficial or deep layers. Colors indicate the cell type and layer for each cell based on layer-specific dissections. (**K**) tSNE visualization of human PV and SST neurons that were putatively located in superficial or deep layers. (**L**,**M**) Global mCH of mouse (L) and human (M) PV and SST neurons. (**N**,**O**) mCH level of mouse (N) and human (O) genes showing layer-specific mCH in PV and SST neurons. The z-score is defined as the mCH value minus its mean over all cells, divided by the standard deviation across cells.

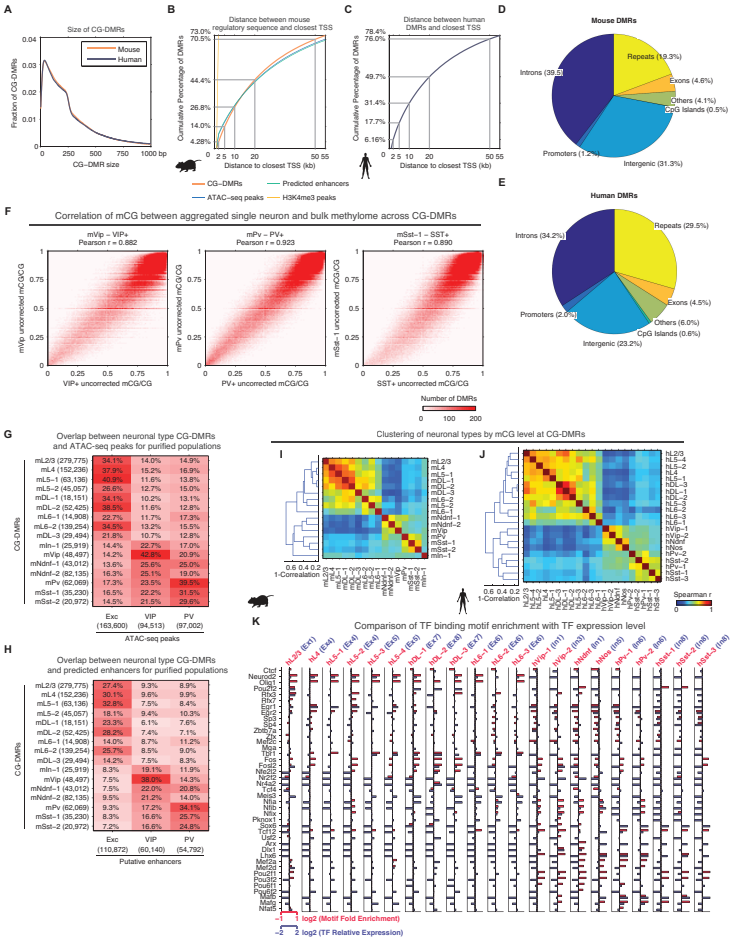

Figure C.15: Neuron-type-specific CG-DMRs reveal regulatory diversity in human and mouse brains. (**A**) Distribution of CG-DMR size. Note that the DMR calling software (methylpy) merges CG positions spaced closer than 250 bp to call DMRs, which accounts for the drop in the frequency of DMRs around 250 bp. (**B**) Distance between closest TSS and mouse regulatory sequences defined by CG-DMRs, enhancers predicted by Regulatory Element Prediction based on Tissue-specific Local Epigenetic marks (REPTILE, (20)), ATAC-seq peaks and H3K4me3 peaks. The curve shows the cumulative percentage of DMRs within a certain distance to the closest TSS. (**C**) Distance between human CG-DMRs and closest TSS. (**D**-**E**) Distribution of mouse (D) and human (E) CG-DMRs in genomic compartments. (**F**) Consistent mCG across CG-DMRs between bulk methylome and aggregated single neuron methylomes. Bulk methylome generated from purified mouse neuronal populations (VIP+, PV+ and SST+) were compared with aggregated single neurons methylomes of corresponding clusters (mVip, mPv and mSst-1). (**G**) Overlap between neuron-type-specific CG-DMRs and ATAC-seq peaks identified from purified neuronal populations. (**H**) Overlap between neuron-type-specific CG-DMRs and putative enhancers predicted from purified neuronal populations. For (G) and (H), the percentage of row features overlapping with column features was shown. **I**-**J**) Hierarchical clustering of neuron clusters by mCG level at CG-DMRs. (**K**) Comparison of TF binding motif enrichment with TF expression level across human neuron clusters. Median TF expression of the best matching snRNA-seq cluster (indicated on the top) identified in (4) for each snmC-seq clusters was shown.

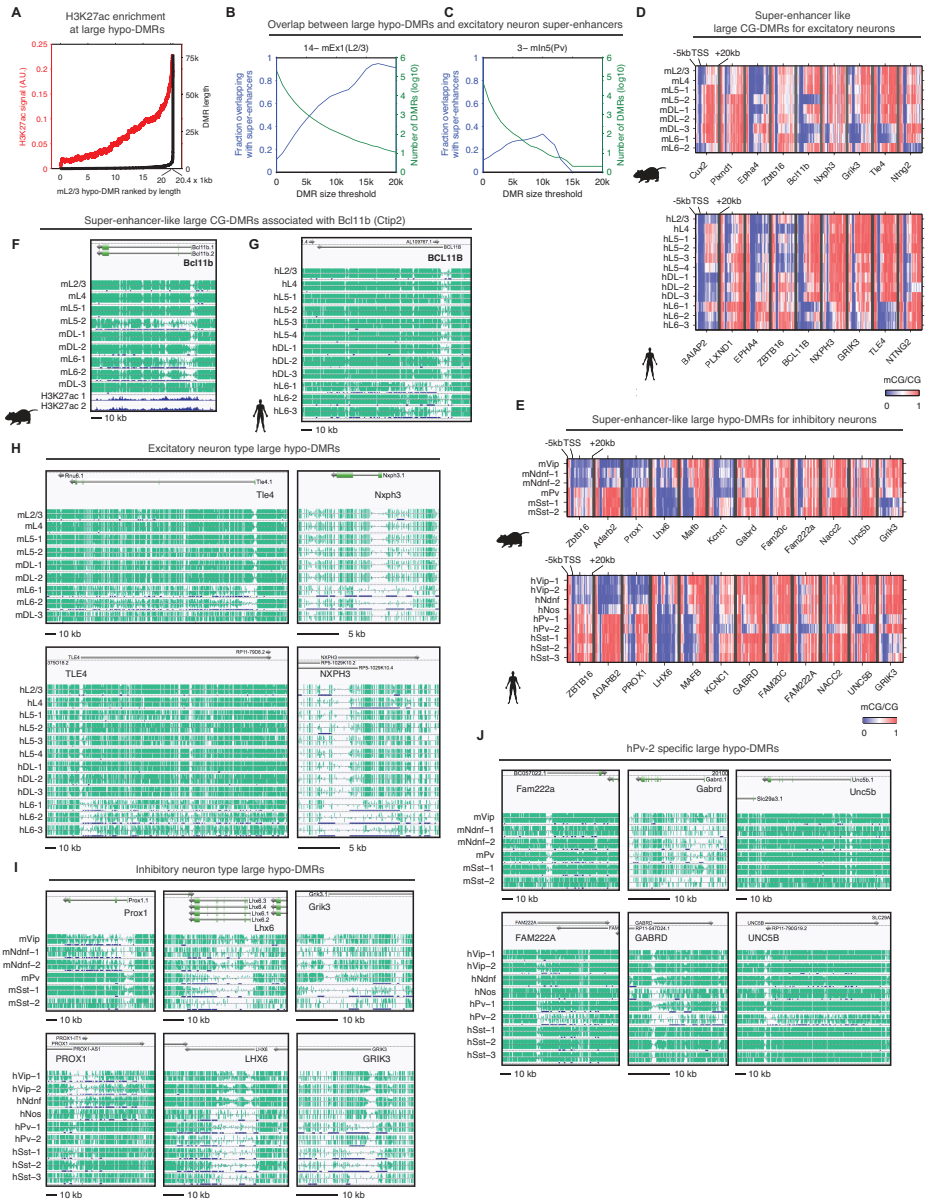

Figure C.16: Identification of neuron type specific large CG-DMRs with super-enhancer like properties. (**A**) Average H3K27ac signal was plotted as a function of CG-DMR size. (**B** and **C**) The fraction of large CG-DMRs overlapping with putative super-enhancers was examined for different DMR size thresholds for identifying large CG-DMRs (blue line). Green line indicates the number of large CG-DMRs found with each DMR size threshold. The overlap between excitatory neuron (Camk2a+) super-enhancers and Layer 2/3 excitatory neuron and PV+ inhibitory neuron large CG-DMRs was shown in (B) and (C), respectively. (**D** and **E**) mCG levels near TSS for super-enhancer-like DMRs in excitatory (D) and inhibitory (E) neurons in mouse and human. (**F**,**G**) Large gene body CG-DMRs and H3K27ac ChIP-seq signal from mouse excitatory neurons at Bcl11b (Ctip2) locus in mouse (F) and human (G). The height of green ticks represents mC level at CG dinucleotides. (**H**-**J**) Browser view of large CG-DMRs for excitatory neurons (H), inhibitory neurons (I) and hPv-2 (J). For F-J, the height of green ticks represents mC level at CG dinucleotides.

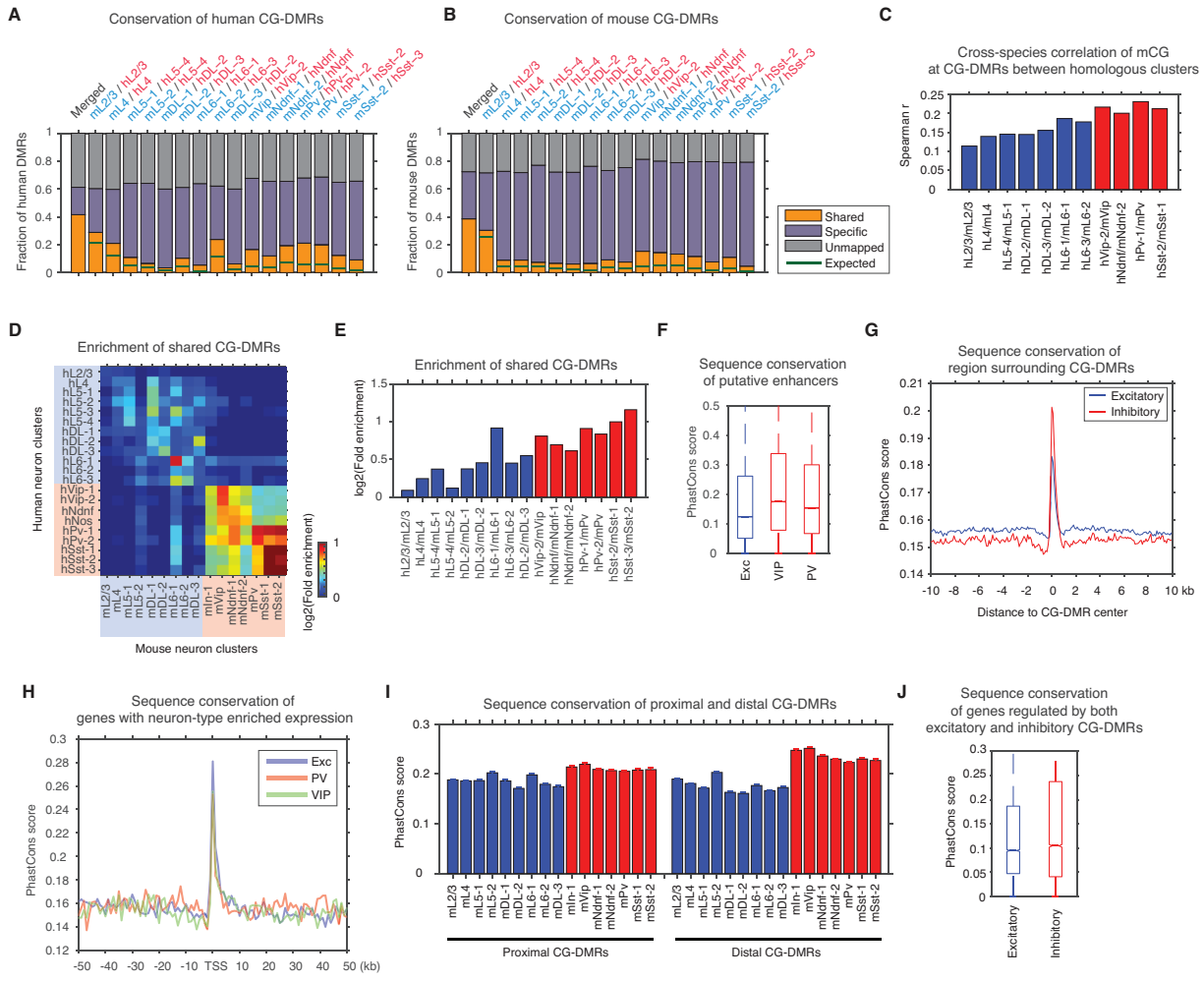

Figure C.17: Regulatory conservation of neuron types. (**A** and **B**) Fractions of human (A) and mouse (B) CG-DMRs that overlapped with CG-DMR of the homologous cluster in the other species (shared) had no overlap with CG-DMRs of the homologous cluster in other species (specific), or had no sequence homology in other species (unmapped). (**C**) Cross-species Spearman correlation of mCG at CG-DMRs between homologous clusters. (**D**-**E**) Enrichment of shared DMRs between all human and mouse clusters (D) and for homologous clusters (E). (**F**-**J**) Sequence conservation at putative enhancers predicted from purified neuronal populations (F), regions surrounding CG-DMRs identified in excitatory and inhibitory neuron clusters (G), genes with preferential expression in purified neuronal populations (H), proximal and distal mouse CG-DMRs (I), and excitatory and inhibitory CG-DMRs associated with the same set of genes (J).

Bibliography

[1] K. Lee, "Little liars: Development of verbal deception in children," *Child Dev. Perspect.*, vol. 7, pp. 91–96, June 2013.

[2] C. Dupont, D. R. Armant, and C. A. Brenner, "Epigenetics: definition, mechanisms and clinical perspective," *Semin. Reprod. Med.*, vol. 27, pp. 351–357, Sept. 2009.

[3] H. Stroud, S. C. Su, S. Hrvatin, A. W. Greben, W. Renthal, L. D. Boxer, M. A. Nagy, D. R. Hochbaum, B. Kinde, H. W. Gabel, and M. E. Greenberg, "Early-Life gene expression in neurons modulates lasting epigenetic states," *Cell*, vol. 171, pp. 1151–1164.e16, Nov. 2017.

[4] R. Lister, E. A. Mukamel, J. R. Nery, M. Urich, C. A. Puddifoot, N. D. Johnson, J. Lucero, Y. Huang, A. J. Dwork, M. D. Schultz, M. Yu, J. Tonti-Filippini, H. Heyn, S. Hu, J. C. Wu, A. Rao, M. Esteller, C. He, F. G. Haghighi, T. J. Sejnowski, M. M. Behrens, and J. R. Ecker, "Global epigenomic reconfiguration during mammalian brain development," *Science*, vol. 341, p. 1237905, Aug. 2013.

[5] A. Mo, E. A. Mukamel, F. P. Davis, C. Luo, G. L. Henry, S. Picard, M. A. Urich, J. R. Nery, T. J. Sejnowski, R. Lister, S. R. Eddy, J. R. Ecker, and J. Nathans, "Epigenomic signatures of neuronal diversity in the mammalian brain," *Neuron*, vol. 86, pp. 1369–1384, June 2015.

[6] I. C. G. Weaver, N. Cervoni, F. A. Champagne, A. C. D'Alessio, S. Sharma, J. R. Seckl, S. Dymov, M. Szyf, and M. J. Meaney, "Epigenetic programming by maternal behavior," *Nat. Neurosci.*, vol. 7, pp. 847–854, Aug. 2004.

[7] G. Turecki and M. J. Meaney, "Effects of the social environment and stress on glucocorticoid receptor gene methylation: A systematic review," *Biol. Psychiatry*, vol. 79, pp. 87–96, Jan. 2016.

[8] T. A. Bedrosian, C. Quayle, N. Novaresi, and F. H. Gage, "Early life experience drives structural variation of neural genomes in mice," *Science*, vol. 359, pp. 1395–1399, Mar. 2018.

[9] W. Xie, C. L. Barr, A. Kim, F. Yue, A. Y. Lee, J. Eubanks, E. L. Dempster, and B. Ren, "Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome," *Cell*, vol. 148, pp. 816–831, Feb. 2012.

[10] 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*, vol. 462, pp. 315–322, Nov. 2009.

[11] 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*, vol. 471, pp. 68–73, Mar. 2011.

[12] F. A. Champagne, "Social and behavioral epigenetics: Evolving perspectives on Nature-Nurture interplay, plasticity, and inheritance," in *The Palgrave Handbook of Biology and Society* (M. Meloni, J. Cromby, D. Fitzgerald, and S. Lloyd, eds.), pp. 227–250, London: Palgrave Macmillan UK, 2018.

[13] M. Weber, J. J. Davies, D. Wittig, E. J. Oakeley, M. Haase, W. L. Lam, and D. Schübeler, "Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells," *Nat. Genet.*, vol. 37, pp. 853–862, Aug. 2005.

[14] P. J. Hurd and C. J. Nelson, "Advantages of next-generation sequencing versus the microarray in epigenetic research," *Brief. Funct. Genomic. Proteomic.*, vol. 8, pp. 174–183, May 2009.

[15] R. Lister and J. R. Ecker, "Finding the fifth base: genome-wide sequencing of cytosine methylation," *Genome Res.*, vol. 19, pp. 959–966, June 2009.

[16] A. Meissner, A. Gnirke, G. W. Bell, B. Ramsahoye, E. S. Lander, and R. Jaenisch, "Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis," *Nucleic Acids Res.*, vol. 33, pp. 5868–5877, Oct. 2005.

[17] E. R. De Kloet, E. Vreugdenhil, M. S. Oitzl, and M. Joëls, "Brain corticosteroid receptor balance in health and disease," *Endocr. Rev.*, vol. 19, pp. 269–301, June 1998.

[18] R. Halder, M. Hennion, R. O. Vidal, O. Shomroni, R.-U. Rahman, A. Rajput, T. P. Centeno, F. van Bebber, V. Capece, J. C. Garcia Vizcaino, A.-L. Schuetz, S. Burkhardt, E. Benito, M. Navarro Sala, S. B. Javan, C. Haass, B. Schmid, A. Fischer, and S. Bonn, "DNA methylation changes in plasticity genes accompany the formation and maintenance of memory," *Nat. Neurosci.*, vol. 19, pp. 102–110, Jan. 2016.

[19] F. D. Heyward and J. D. Sweatt, "DNA methylation in memory formation: Emerging insights," *Neuroscientist*, vol. 21, pp. 475–489, Oct. 2015.

[20] M. L. Lehmann and M. Herkenham, "Environmental enrichment confers stress resiliency to social defeat through an infralimbic cortex-dependent neuroanatomical pathway," *J. Neurosci.*, vol. 31, pp. 6159–6173, Apr. 2011.

[21] G. D. Clemenson, W. Deng, and F. H. Gage, "Environmental enrichment and neurogenesis: from mice to humans," *Current Opinion in Behavioral Sciences*, vol. 4, pp. 56–62, Aug. 2015.

[22] C. Rampon, C. H. Jiang, H. Dong, Y. P. Tang, D. J. Lockhart, P. G. Schultz, J. Z. Tsien, and Y. Hu, "Effects of environmental enrichment on gene expression in the brain," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 97, pp. 12880–12884, Nov. 2000.

[23] P. S. Eriksson, E. Perfilieva, T. Björk-Eriksson, A. M. Alborn, C. Nordborg, D. A. Peterson, and F. H. Gage, "Neurogenesis in the adult human hippocampus," *Nat. Med.*, vol. 4, pp. 1313–1317, Nov. 1998.

[24] M. S. Fanselow and H.-W. Dong, "Are the dorsal and ventral hippocampus functionally distinct structures?," *Neuron*, vol. 65, pp. 7–19, Jan. 2010.

[25] M. F. Lyon, "Gene action in the x-chromosome of the mouse (mus musculus l.)," 1961.

[26] B. Reinius, C. Shi, L. Hengshuo, K. S. Sandhu, K. J. Radomska, G. D. Rosen, L. Lu, K. Kullander, R. W. Williams, and E. Jazin, "Female-biased expression of long non-coding RNAs in domains that escape x-inactivation in mouse," *BMC Genomics*, vol. 11, p. 614, Nov. 2010.

[27] L. Carrel and H. F. Willard, "X-inactivation profile reveals extensive variability in x-linked gene expression in females," *Nature*, vol. 434, pp. 400–404, Mar. 2005.

[28] J. F. Rovet, "The psychoeducational characteristics of children with turner syndrome," *J. Learn. Disabil.*, vol. 26, pp. 333–341, May 1993.

[29] A. J. Sharp, E. Stathaki, E. Migliavacca, M. Brahmachary, S. B. Montgomery, Y. Dupre, and S. E. Antonarakis, "DNA methylation profiles of human active and inactive X chromosomes," *Genome Res.*, vol. 21, pp. 1592–1600, Oct. 2011.

[30] Y. Hoki, N. Kimura, M. Kanbayashi, Y. Amakawa, T. Ohhata, H. Sasaki, and T. Sado, "A proximal conserved repeat in the xist gene is essential as a genomic element for x-inactivation in mouse," *Development*, vol. 136, pp. 139–146, Jan. 2009.

[31] E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, and S. A. McCarroll, "Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets," *Cell*, vol. 161, pp. 1202–1214, May 2015.

[32] V. Proserpio and T. Lönnberg, "Single-cell technologies are revolutionizing the approach to rare cells," *Immunol. Cell Biol.*, vol. 94, pp. 225–229, Mar. 2016.

[33] A. A. Kolodziejczyk, J. K. Kim, V. Svensson, J. C. Marioni, and S. A. Teichmann, "The technology and biology of single-cell RNA sequencing," *Mol. Cell*, vol. 58, pp. 610–620, May 2015.

[34] B. Lacar, S. B. Linker, B. N. Jaeger, S. Krishnaswami, J. Barron, M. Kelder, S. Parylak, A. Paquola, P. Venepally, M. Novotny, C. O'Connor, C. Fitzpatrick, J. Erwin, J. Y. Hsu, D. Husband, M. J. McConnell, R. Lasken, and F. H. Gage, "Nuclear RNA-seq of single neurons reveals molecular signatures of activation," *Nat. Commun.*, vol. 7, p. 11022, Apr. 2016.

[35] A. Kepecs and G. Fishell, "Interneuron cell types are fit to function," *Nature*, vol. 505, pp. 318–326, Jan. 2014.

[36] A. Zeisel, A. B. Muñoz-Manchado, S. Codeluppi, P. Lönnerberg, G. La Manno, A. Juréus, S. Marques, H. Munguba, L. He, C. Betsholtz, C. Rolny, G. Castelo-Branco, J. Hjerling-Leffler, and S. Linnarsson, "Brain structure. cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq," *Science*, vol. 347, pp. 1138–1142, Mar. 2015.

[37] B. Tasic, V. Menon, T. N. Nguyen, T. K. Kim, T. Jarsky, Z. Yao, B. Levi, L. T. Gray, S. A. Sorensen, T. Dolbeare, D. Bertagnolli, J. Goldy, N. Shapovalova, S. Parry, C. Lee, K. Smith, A. Bernard, L. Madisen, S. M. Sunkin, M. Hawrylycz, C. Koch, and H. Zeng, "Adult mouse cortical cell taxonomy revealed by single cell transcriptomics," *Nat. Neurosci.*, vol. 19, pp. 335–346, Feb. 2016.

[38] B. Tasic, Z. Yao, K. A. Smith, L. Graybuck, T. N. Nguyen, and others, "Shared and distinct transcriptomic cell types across neocortical areas," *bioRxiv*, 2017.

[39] B. B. Lake, R. Ai, G. E. Kaeser, N. S. Salathia, Y. C. Yung, R. Liu, A. Wildberg, D. Gao, H.-L. Fung, S. Chen, R. Vijayaraghavan, J. Wong, A. Chen, X. Sheng, F. Kaper, R. Shen, M. Ronaghi, J.-B. Fan, W. Wang, J. Chun, and K. Zhang, "Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain," *Science*, vol. 352, pp. 1586–1590, June 2016.

[40] M. S. Cembrowski, L. Wang, K. Sugino, B. C. Shields, and N. Spruston, "Hipposeq: a comprehensive RNA-seq database of gene expression in hippocampal principal neurons," *Elife*, vol. 5, Apr. 2016.

[41] J. P. Buschdorf and M. J. Meaney, "Epigenetics/Programming in the HPA axis," *Compr. Physiol.*, vol. 6, pp. 87–110, Dec. 2015.

[42] B. Yao, K. M. Christian, C. He, P. Jin, G.-L. Ming, and H. Song, "Epigenetic mechanisms in neurogenesis," *Nat. Rev. Neurosci.*, vol. 17, pp. 537–549, Sept. 2016.

[43] J. J. Day, D. Childs, M. C. Guzman-Karlsson, M. Kibe, J. Moulden, E. Song, A. Tahir, and J. D. Sweatt, "DNA methylation regulates associative reward learning," *Nat. Neurosci.*, vol. 16, pp. 1445–1452, Oct. 2013.

[44] I. B. Zovkic, M. C. Guzman-Karlsson, and J. D. Sweatt, "Epigenetic regulation of memory formation and maintenance," *Learn. Mem.*, vol. 20, pp. 61–74, Jan. 2013.

[45] T.-Y. Zhang and M. J. Meaney, "Epigenetics and the environmental regulation of the genome and its function," *Annu. Rev. Psychol.*, vol. 61, pp. 439–466, Dec. 2009.

[46] J. Cholewa-Waclaw, A. Bird, M. von Schimmelmann, A. Schaefer, H. Yu, H. Song, R. Madabhushi, and L.-H. Tsai, "The role of epigenetic mechanisms in the regulation of gene expression in the nervous system," *J. Neurosci.*, vol. 36, pp. 11427–11434, Nov. 2016.

[47] Z. D. Smith and A. Meissner, "DNA methylation: roles in mammalian development," *Nat. Rev. Genet.*, vol. 14, pp. 204–220, Mar. 2013.

[48] J. Scholz, R. Allemang-Grand, J. Dazai, and J. P. Lerch, "Environmental enrichment is associated with rapid volumetric brain changes in adult mice," *Neuroimage*, vol. 109, pp. 190–198, Apr. 2015.

[49] S. Rizzi, P. Bianchi, S. Guidi, E. Ciani, and R. Bartesaghi, "Impact of environmental enrichment on neurogenesis in the dentate gyrus during the early postnatal period," *Brain Res.*, vol. 1415, pp. 23–33, Sept. 2011.

[50] R. S. Nowakowski, S. B. Lewin, and M. W. Miller, "Bromodeoxyuridine immunohistochemical determination of the lengths of the cell cycle and the DNA-synthetic phase for an anatomically defined population," *J. Neurocytol.*, vol. 18, pp. 311–318, June 1989.

[51] N. Habib, Y. Li, M. Heidenreich, L. Swiech, I. Avraham-Davidi, J. J. Trombetta, C. Hession, F. Zhang, and A. Regev, "Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons," *Science*, vol. 353, pp. 925–928, Aug. 2016.

[52] Y. Zhang, K. Chen, S. A. Sloan, M. L. Bennett, A. R. Scholze, S. O'Keeffe, H. P. Phatnani, P. Guarnieri, C. Caneda, N. Ruderisch, S. Deng, S. A. Liddelow, C. Zhang, R. Daneman, T. Maniatis, B. A. Barres, and J. Q. Wu, "An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex," *J. Neurosci.*, vol. 34, pp. 11929–11947, Sept. 2014.

[53] R. Beckervordersandforth, C.-L. Zhang, and D. C. Lie, "Transcription-Factor-Dependent control of adult hippocampal neurogenesis," *Cold Spring Harb. Perspect. Biol.*, vol. 7, p. a018879, Oct. 2015.

[54] J. S. Snyder, R. Radik, J. M. Wojtowicz, and H. A. Cameron, "Anatomical gradients of adult neurogenesis and activity: young neurons in the ventral dentate gyrus are activated by water maze training," *Hippocampus*, vol. 19, pp. 360–370, Apr. 2009.

[55] D. K. Ma, M.-H. Jang, J. U. Guo, Y. Kitabatake, M.-L. Chang, N. Pow-Anpongkul, R. A. Flavell, B. Lu, G.-L. Ming, and H. Song, "Neuronal activity-induced gadd45b promotes epigenetic DNA demethylation and adult neurogenesis," *Science*, vol. 323, pp. 1074–1077, Feb. 2009.

[56] S. Kriaucionis and N. Heintz, "The nuclear DNA base 5-hydroxymethylcytosine is present in purkinje neurons and the brain," *Science*, vol. 324, pp. 929–930, May 2009.

[57] M. Yu, G. C. Hon, K. E. Szulwach, C.-X. Song, P. Jin, B. Ren, and C. He, "Tet-assisted bisulfite sequencing of 5-hydroxymethylcytosine," *Nat. Protoc.*, vol. 7, pp. 2159–2170, Dec. 2012.

[58] P. Fuentealba, T. Klausberger, T. Karayannis, W. Y. Suen, J. Huck, R. Tomioka, K. Rockland, M. Capogna, M. Studer, M. Morales, and P. Somogyi, "Expression of COUP-TFII nuclear receptor in restricted GABAergic neuronal populations in the adult rat hippocampus," *J. Neurosci.*, vol. 30, pp. 1595–1609, Feb. 2010.

[59] 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.*, vol. 17, pp. 215–222, Feb. 2014.

[60] M. Yu, G. C. Hon, K. E. Szulwach, C.-X. Song, L. Zhang, A. Kim, X. Li, Q. Dai, Y. Shen, B. Park, J.-H. Min, P. Jin, B. Ren, and C. He, "Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome," *Cell*, vol. 149, pp. 1368–1380, June 2012.

[61] H. Feng, K. N. Conneely, and H. Wu, "A bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data," *Nucleic Acids Res.*, vol. 42, p. e69, Apr. 2014.

[62] Y. Hara, T. Nomura, K. Yoshizaki, J. Frisén, and N. Osumi, "Impaired hippocampal neurogenesis and vascular formation in ephrin?a5?deficient mice," *Stem Cells*, 2010.

[63] W. Kang and J. M. Hébert, "FGF signaling is necessary for neurogenesis in young mice and sufficient to reverse its decline in old mice," *J. Neurosci.*, vol. 35, pp. 10217–10223, July 2015.

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

[65] Z. Gao, K. Ure, J. L. Ables, D. C. Lagace, K.-A. Nave, S. Goebbels, A. J. Eisch, and J. Hsieh, "Neurod1 is essential for the survival and maturation of adult-born neurons," *Nat. Neurosci.*, vol. 12, pp. 1090–1092, Sept. 2009.

[66] T. Kuwabara, J. Hsieh, A. Muotri, G. Yeo, M. Warashina, D. C. Lie, L. Moore, K. Nakashima, M. Asashima, and F. H. Gage, "Wnt-mediated activation of NeuroD1 and retro-elements during adult neurogenesis," *Nat. Neurosci.*, vol. 12, pp. 1097–1105, Sept. 2009.

[67] A. Pataskar, J. Jung, P. Smialowski, F. Noack, F. Calegari, T. Straub, and V. K. Tiwari, "NeuroD1 reprograms chromatin and transcription factor landscapes to induce the neuronal program," *EMBO J.*, vol. 35, pp. 24–45, Jan. 2016.

[68] H. Li, J. C. Radford, M. J. Ragusa, K. L. Shea, S. R. McKercher, J. D. Zaremba, W. Soussou, Z. Nie, Y.-J. Kang, N. Nakanishi, S.-I. Okamoto, A. J. Roberts, J. J. Schwarz, and S. A. Lipton, "Transcription factor MEF2C influences neural stem/progenitor cell differentiation and maturation in vivo," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, pp. 9397–9402, July 2008.

[69] C. Menke, M. Cionni, T. Siggers, M. L. Bulyk, D. R. Beier, and R. W. Stottmann, "Grhl2 is required in nonneural tissues for neural progenitor survival and forebrain development," *Genesis*, July 2015.

[70] H. Van Praag, G. Kempermann, and F. H. Gage, "Neural consequences of enviromental enrichment," *Nat. Rev. Neurosci.*, vol. 1, no. 3, pp. 191–198, 2000.

[71] A. Mansouri and P. Gruss, "Pax3 and pax7 are expressed in commissural neurons and restrict ventral neuronal identity in the spinal cord," *Mech. Dev.*, vol. 78, pp. 171–178, Nov. 1998.

[72] M. S. Cembrowski, J. L. Bachman, L. Wang, K. Sugino, B. C. Shields, and N. Spruston, "Spatial Gene-Expression gradients underlie prominent heterogeneity of CA1 pyramidal neurons," *Neuron*, vol. 89, pp. 351–368, Jan. 2016.

[73] N. Kuzumaki, D. Ikegami, R. Tamura, N. Hareyama, S. Imai, M. Narita, K. Torigoe, K. Niikura, H. Takeshima, T. Ando, K. Igarashi, J. Kanno, T. Ushijima, T. Suzuki, and M. Narita, "Hippocampal epigenetic modification at the brain-derived neurotrophic factor gene induced by an enriched environment," *Hippocampus*, vol. 21, pp. 127–132, Feb. 2011.

[74] C. Rossi, A. Angelucci, L. Costantin, C. Braschi, M. Mazzantini, F. Babbini, M. E. Fabbri, L. Tessarollo, L. Maffei, N. Berardi, and Others, "Brain-derived neurotrophic factor (BDNF) is required for the enhancement of hippocampal neurogenesis following environmental enrichment," *Eur. J. Neurosci.*, vol. 24, no. 7, pp. 1850–1856, 2006.

[75] L. Cao, X. Jiao, D. S. Zuzga, Y. Liu, D. M. Fong, D. Young, and M. J. During, "VEGF links hippocampal activity with neurogenesis, learning and memory," *Nat. Genet.*, vol. 36, pp. 827–835, Aug. 2004.

[76] G. Gangarossa, S. Longueville, D. De Bundel, J. Perroy, D. Hervé, J.-A. Girault, and E. Valjent, "Characterization of dopamine D1 and D2 receptor-expressing neurons in the mouse hippocampus," *Hippocampus*, vol. 22, pp. 2199–2207, Dec. 2012.

[77] N. Lemon and D. Manahan-Vaughan, "Dopamine D1/D5 receptors gate the acquisition of novel information through hippocampal long-term potentiation and long-term depression," *J. Neurosci.*, vol. 26, pp. 7723–7729, July 2006.

[78] G. Segovia, A. del Arco, and F. Mora, "Environmental enrichment, prefrontal cortex, stress, and aging of the brain," *J. Neural Transm.*, vol. 116, pp. 1007–1016, Aug. 2009.

[79] M. Tahiliani, K. P. Koh, Y. Shen, W. A. Pastor, H. Bandukwala, Y. Brudno, S. Agarwal, L. M. Iyer, D. R. Liu, L. Aravind, and A. Rao, "Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1," *Science*, vol. 324, pp. 930–935, May 2009.

[80] M. Banasr, A. Soumier, M. Hery, E. Mocaër, and A. Daszuta, "Agomelatine, a new antidepressant, induces regional changes in hippocampal neurogenesis," *Biol. Psychiatry*, vol. 59, pp. 1087–1096, June 2006.

[81] C. Angermueller, S. J. Clark, H. J. Lee, I. C. Macaulay, M. J. Teng, T. X. Hu, F. Krueger, S. A. Smallwood, C. P. Ponting, T. Voet, G. Kelsey, O. Stegle, and W. Reik, "Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity," *Nat. Methods*, vol. 13, pp. 229–232, Mar. 2016.

[82] C. Luo, C. L. Keown, L. Kurihara, J. Zhou, Y. He, J. Li, R. Castanon, J. Lucero, J. R. Nery, J. P. Sandoval, B. Bui, T. J. Sejnowski, T. T. Harkins, E. A. Mukamel, M. M. Behrens, and J. R. Ecker, "Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex," *Science*, vol. 357, pp. 600–604, Aug. 2017.

[83] L. S. Cahill, C. L. Laliberté, J. Ellegood, S. Spring, J. A. Gleave, M. C. v. Eede, J. P. Lerch, and R. M. Henkelman, "Preparation of fixed mouse brains for MRI," *Neuroimage*, vol. 60, pp. 933–939, Apr. 2012.

[84] J. P. Lerch, J. G. Sled, and R. M. Henkelman, "MRI phenotyping of genetically altered mice," *Methods Mol. Biol.*, vol. 711, pp. 349–361, 2011.

[85] J. Dazai, S. Spring, L. S. Cahill, and R. M. Henkelman, "Multiple-mouse neuroanatomical magnetic resonance imaging," *J. Vis. Exp.*, Feb. 2011.

[86] J. P. Lerch, J. B. Carroll, S. Spring, L. N. Bertram, C. Schwab, M. R. Hayden, and R. M. Henkelman, "Automated deformation analysis in the YAC128 huntington disease mouse model," *Neuroimage*, vol. 39, pp. 32–39, Jan. 2008.

[87] D. L. Collins, P. Neelin, T. M. Peters, and A. C. Evans, "Automatic 3D intersubject registration of MR volumetric data in standardized talairach space," *J. Comput. Assist. Tomogr.*, vol. 18, pp. 192–205, Mar. 1994.

[88] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee, "Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain," *Med. Image Anal.*, vol. 12, pp. 26–41, Feb. 2008.

[89] M. Friedel, M. C. van Eede, J. Pipitone, M. M. Chakravarty, and J. P. Lerch, "Pydpiper: a flexible toolkit for constructing novel registration pipelines," *Front. Neuroinform.*, vol. 8, p. 67, July 2014.

[90] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmertest: tests in linear mixed effects models. R package version 2.0-20," 2015.

[91] G. Paxinos and K. B. J. Franklin, *The Mouse Brain in Stereotaxic Coordinates*. Gulf Professional Publishing, 2004.

[92] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, "STAR: ultrafast universal RNA-seq aligner," *Bioinformatics*, vol. 29, pp. 15–21, Jan. 2013.

[93] D. J. McCarthy, Y. Chen, and G. K. Smyth, "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation," *Nucleic Acids Res.*, vol. 40, pp. 4288–4297, May 2012.

[94] H. Wu, C. Wang, and Z. Wu, "A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data," *Biostatistics*, vol. 14, pp. 232–243, Apr. 2013.

[95] W. Reik and J. Walter, "Genomic imprinting: parental influence on the genome," *Nat. Rev. Genet.*, vol. 2, pp. 21–32, Jan. 2001.

[96] J. McGrath and D. Solter, "Completion of mouse embryogenesis requires both the maternal and paternal genomes," *Cell*, vol. 37, pp. 179–183, May 1984.

[97] K. Buiting, S. Saitoh, S. Gross, B. Dittrich, S. Schwartz, R. D. Nicholls, and B. Horsthemke, "Inherited microdeletions in the angelman and Prader–Willi syndromes define an imprinting centre on human chromosome 15," *Nat. Genet.*, vol. 9, pp. 395–400, Apr. 1995.

[98] S. Augui, E. P. Nora, and E. Heard, "Regulation of x-chromosome inactivation by the x-inactivation centre," *Nat. Rev. Genet.*, vol. 12, pp. 429–442, June 2011.

[99] J. B. Berletch, W. Ma, F. Yang, J. Shendure, W. S. Noble, C. M. Disteche, and X. Deng, "Escape from X inactivation varies in mouse tissues," *PLoS Genet.*, vol. 11, p. e1005079, Mar. 2015.

[100] 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.*, vol. 112, pp. 5509–5514, Apr. 2015.

[101] H. W. Gabel, B. Kinde, H. Stroud, C. S. Gilbert, D. A. Harmin, N. R. Kastan, M. Hemberg, D. H. Ebert, and M. E. Greenberg, "Disruption of DNA-methylation-dependent long gene repression in rett syndrome," *Nature*, vol. 522, pp. 89–93, June 2015.

[102] 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*, vol. 157, pp. 979–991, May 2014.

[103] 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*, vol. 523, pp. 212–216, July 2015.

[104] J. T. Lee, W. M. Strauss, J. A. Dausman, and R. Jaenisch, "A 450 kb transgene displays properties of the mammalian x-inactivation center," *Cell*, vol. 86, pp. 83–94, July 1996.

[105] E. P. Nora, B. R. Lajoie, E. G. Schulz, L. Giorgetti, I. Okamoto, N. Servant, T. Piolot, N. L. van Berkum, J. Meisig, J. Sedat, J. Gribnau, E. Barillot, N. Blüthgen, J. Dekker, and E. Heard, "Spatial partitioning of the regulatory landscape of the x-inactivation centre," *Nature*, vol. 485, pp. 381–385, May 2012.

[106] B. Reinius, M. M. Johansson, K. J. Radomska, E. H. Morrow, G. K. Pandey, C. Kanduri, R. Sandberg, R. W. Williams, and E. Jazin, "Abundance of female-biased and paucity of male-biased somatically expressed genes on the mouse x-chromosome," *BMC Genomics*, vol. 13, no. 1, pp. 1–18, 2012.

[107] A. H. Horakova, J. M. Calabrese, C. R. McLaughlin, D. C. Tremblay, T. Magnuson, and B. P. Chadwick, "The mouse DXZ4 homolog retains ctcf binding and proximity to pls3 despite substantial organizational differences compared to the primate macrosatellite," *Genome Biol.*, vol. 13, p. R70, Aug. 2012.

[108] X. Deng, W. Ma, V. Ramani, A. Hill, F. Yang, F. Ay, J. B. Berletch, C. A. Blau, J. Shendure, Z. Duan, W. S. Noble, and C. M. Disteche, "Bipartite structure of the inactive mouse X chromosome," *Genome Biol.*, vol. 16, p. 152, Aug. 2015.

[109] L. Burger, D. Gaidatzis, D. Schübeler, and M. B. Stadler, "Identification of active regulatory regions from DNA methylation data," *Nucleic Acids Res.*, vol. 41, p. e155, Sept. 2013.

[110] J. E. Joo, B. Novakovic, M. Cruickshank, L. W. Doyle, J. M. Craig, and R. Saffery, "Human active x-specific DNA methylation events showing stability across time and tissues," *Eur. J. Hum. Genet.*, vol. 22, pp. 1376–1381, Apr. 2014.

[111] J. A. Wamstad, C. M. Corcoran, A. M. Keating, and V. J. Bardwell, "Role of the transcriptional corepressor bcor in embryonic stem cell differentiation and early embryonic development," *PLoS One*, vol. 3, p. e2814, July 2008.

[112] A. M. Cotton, E. M. Price, M. J. Jones, B. P. Balaton, M. S. Kobor, and C. J. Brown, "Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and x-chromosome inactivation," *Hum. Mol. Genet.*, vol. 24, pp. 1528–1539, Mar. 2015.

[113] B. Kinde, H. W. Gabel, C. S. Gilbert, E. C. Griffith, and M. E. Greenberg, "Reading the unique DNA methylation landscape of the brain: Non-CpG methylation, hydroxymethylation, and MeCP2," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 112, pp. 6800–6806, June 2015.

[114] G. C. Hon, N. Rajagopal, Y. Shen, D. F. McCleary, F. Yue, M. D. Dang, and B. Ren, "Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues," *Nat. Genet.*, vol. 45, pp. 1198–1206, Oct. 2013.

[115] P. A. Lingenfelter, D. A. Adler, D. Poslinski, S. Thomas, R. W. Elliott, V. M. Chapman, and C. M. Disteche, "Escape from X inactivation of smcx is preceded by silencing during mouse development," *Nat. Genet.*, vol. 18, pp. 212–213, Mar. 1998.

[116] B. M. Nugent, C. L. Wright, A. C. Shetty, G. E. Hodes, K. M. Lenz, A. Mahurkar, S. J. Russo, S. E. Devine, and M. M. McCarthy, "Brain feminization requires active repression of masculinization via DNA methylation," *Nat. Neurosci.*, vol. 18, pp. 690–697, May 2015.

[117] M. M. McCarthy, A. P. Auger, T. L. Bale, G. J. De Vries, G. A. Dunn, N. G. Forger, E. K. Murray, B. M. Nugent, J. M. Schwarz, and M. E. Wilson, "The epigenetics of sex differences in the brain," *J. Neurosci.*, vol. 29, pp. 12815–12823, Oct. 2009.

[118] J. A. Clayton and F. S. Collins, "Policy: NIH to balance sex in cell and animal studies," *Nature*, vol. 509, pp. 282–283, May 2014.

[119] F. Yang, T. Babak, J. Shendure, and C. M. Disteche, "Global survey of escape from X inactivation by RNA-sequencing in mouse," *Genome Res.*, vol. 20, pp. 614–622, May 2010.

[120] T. M. Keane, L. Goodstadt, P. Danecek, M. A. White, K. Wong, B. Yalcin, A. Heger, A. Agam, G. Slater, M. Goodson, N. A. Furlotte, E. Eskin, C. Nellåker, H. Whitley, J. Cleak, D. Janowitz, P. Hernandez-Pliego, A. Edwards, T. G. Belgard, P. L. Oliver, R. E. McIntyre, A. Bhomra, J. Nicod, X. Gan, W. Yuan, L. van der Weyden, C. A. Steward, S. Bala, J. Stalker, R. Mott, R. Durbin, I. J. Jackson, A. Czechanski, J. A. Guerra-Assunção, L. R. Donahue, L. G. Reinholdt, B. A. Payseur, C. P. Ponting, E. Birney, J. Flint, and D. J. Adams, "Mouse genomic variation and its effect on phenotypes and gene regulation," *Nature*, vol. 477, pp. 289–294, Sept. 2011.

[121] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg, "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions," *Genome Biol.*, vol. 14, p. R36, Apr. 2013.

[122] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter, "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation," *Nat. Biotechnol.*, vol. 28, pp. 511–515, May 2010.

[123] B. J. Molyneaux, P. Arlotta, J. R. L. Menezes, and J. D. Macklis, "Neuronal subtype specification in the cerebral cortex," *Nat. Rev. Neurosci.*, vol. 8, pp. 427–437, June 2007.

[124] A. Kozlenkov, M. Wang, P. Roussos, S. Rudchenko, M. Barbu, M. Bibikova, B. Klotzle, A. J. Dwork, B. Zhang, Y. L. Hurd, E. V. Koonin, M. Wegner, and S. Dracheva, "Substantial DNA methylation differences between two major neuronal subtypes in human brain," *Nucleic Acids Res.*, vol. 44, pp. 2593–2612, Apr. 2016.

[125] A. Mo, C. Luo, F. P. Davis, E. A. Mukamel, G. L. Henry, J. R. Nery, M. A. Urich, S. Picard, R. Lister, S. R. Eddy, M. A. Beer, J. R. Ecker, and J. Nathans, "Epigenomic landscapes of retinal rods and cones," *Elife*, vol. 5, p. e11613, Mar. 2016.

[126] S. A. Smallwood, H. J. Lee, C. Angermueller, F. Krueger, H. Saadeh, J. Peat, S. R. Andrews, O. Stegle, W. Reik, and G. Kelsey, "Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity," *Nat. Methods*, vol. 11, pp. 817–820, Aug. 2014.

[127] M. Farlik, N. C. Sheffield, A. Nuzzo, P. Datlinger, A. Schönegger, J. Klughammer, and C. Bock, "Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics," *Cell Rep.*, vol. 10, pp. 1386–1397, Mar. 2015.

[128] Y. Huang, W. A. Pastor, Y. Shen, M. Tahiliani, D. R. Liu, and A. Rao, "The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing," *PLoS One*, vol. 5, p. e8888, Jan. 2010.

[129] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[130] C. P. Wonders and S. A. Anderson, "The origin and specification of cortical interneurons," *Nat. Rev. Neurosci.*, vol. 7, pp. 687–696, Sept. 2006.

[131] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," *KDD*, 1996.

[132] E. S. Lein, M. J. Hawrylycz, N. Ao, M. Ayres, A. Bensinger, A. Bernard, A. F. Boe, M. S. Boguski, K. S. Brockway, E. J. Byrnes, L. Chen, L. Chen, T.-M. Chen, M. C. Chin, J. Chong, B. E. Crook, A. Czaplinska, C. N. Dang, S. Datta, N. R. Dee, A. L. Desaki, T. Desta, E. Diep, T. A. Dolbeare, M. J. Donelan, H.-W. Dong, J. G. Dougherty, B. J. Duncan, A. J. Ebbert, G. Eichele, L. K. Estin, C. Faber, B. A. Facer, R. Fields, S. R. Fischer, T. P. Fliss, C. Frensley, S. N. Gates, K. J. Glattfelder, K. R. Halverson, M. R. Hart, J. G. Hohmann, M. P. Howell, D. P. Jeung, R. A. Johnson, P. T. Karr, R. Kawal, J. M. Kidney, R. H. Knapik, C. L. Kuan, J. H. Lake, A. R. Laramee, K. D. Larsen, C. Lau, T. A. Lemon, A. J. Liang, Y. Liu, L. T. Luong, J. Michaels, J. J. Morgan, R. J. Morgan, M. T. Mortrud, N. F. Mosqueda, L. L. Ng, R. Ng, G. J. Orta, C. C. Overly, T. H. Pak, S. E. Parry, S. D. Pathak, O. C. Pearson, R. B. Puchalski, Z. L. Riley, H. R. Rockett, S. A. Rowland, J. J. Royall, M. J. Ruiz, N. R. Sarno, K. Schaffnit, N. V. Shapovalova, T. Sivisay, C. R. Slaughterbeck, S. C. Smith, K. A. Smith, B. I. Smith, A. J. Sodt, N. N. Stewart, K.-R. Stumpf, S. M. Sunkin, M. Sutram, A. Tam, C. D. Teemer, C. Thaller, C. L. Thompson,

L. R. Varnam, A. Visel, R. M. Whitlock, P. E. Wohnoutka, C. K. Wolkey, V. Y. Wong, M. Wood, M. B. Yaylaoglu, R. C. Young, B. L. Youngstrom, X. F. Yuan, B. Zhang, T. A. Zwingman, and A. R. Jones, "Genome-wide atlas of gene expression in the adult mouse brain," *Nature*, vol. 445, pp. 168–176, Jan. 2007.

[133] S. A. Sorensen, A. Bernard, V. Menon, J. J. Royall, K. J. Glattfelder, T. Desta, K. Hirokawa, M. Mortrud, J. A. Miller, H. Zeng, J. G. Hohmann, A. R. Jones, and E. S. Lein, "Correlated gene expression and target specificity demonstrate excitatory projection neuron diversity," *Cereb. Cortex*, vol. 25, pp. 433–449, Feb. 2015.

[134] 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 Mouse ENCODE Consortium, "A comparative encyclopedia of DNA elements in the mouse genome," *Nature*, vol. 515, pp. 355–364, Nov. 2014.

[135] Y. He, D. U. Gorkin, D. E. Dickel, J. R. Nery, R. G. Castanon, A. Y. Lee, Y. Shen, A. Visel, L. A. Pennacchio, B. Ren, and J. R. Ecker, "Improved regulatory element prediction based on tissue-specific local epigenomic signatures," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 114, pp. E1633–E1640, Feb. 2017.

[136] A. B. Stergachis, S. Neph, R. Sandstrom, E. Haugen, A. P. Reynolds, M. Zhang, R. Byron, T. Canfield, S. Stelhing-Sun, K. Lee, R. E. Thurman, S. Vong, D. Bates, F. Neri, M. Diegel, E. Giste, D. Dunn, J. Vierstra, R. S. Hansen, A. K. Johnson, P. J. Sabo, M. S. Wilken, T. A. Reh, P. M. Treuting, R. Kaul, M. Groudine, M. A. Bender, E. Borenstein, and J. A. Stamatoyannopoulos, "Conservation of trans-acting circuitry during mammalian regulatory evolution," *Nature*, vol. 515, pp. 365–370, Nov. 2014.

[137] W. A. Whyte, D. A. Orlando, D. Hnisz, B. J. Abraham, C. Y. Lin, M. H. Kagey, P. B. Rahl, T. I. Lee, and R. A. Young, "Master transcription factors and mediator establish super-enhancers at key cell identity genes," *Cell*, vol. 153, pp. 307–319, Apr. 2013.

[138] D. V. Hansen, J. H. Lui, P. R. L. Parker, and A. R. Kriegstein, "Neurogenic radial glia in the outer subventricular zone of human neocortex," *Nature*, vol. 464, pp. 554–561, Mar. 2010.

[139] A. A. Pollen, T. J. Nowakowski, J. Chen, H. Retallack, C. Sandoval-Espinosa, C. R. Nicholas, J. Shuga, S. J. Liu, M. C. Oldham, A. Diaz, D. A. Lim, A. A. Leyrat, J. A. West, and A. R. Kriegstein, "Molecular identity of human outer radial glia during cortical development," *Cell*, vol. 163, pp. 55–67, Sept. 2015.

[140] D. Karolchik, A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, D. Haussler, and W. J. Kent, "The UCSC table browser data retrieval tool," *Nucleic Acids Res.*, vol. 32, pp. D493–6, Jan. 2004.

[141] C. M. Williamson, J. A. Skinner, G. Kelsey, and J. Peters, "Alternative non-coding splice variants of nespas, an imprinted gene antisense to nesp in the gnas imprinting cluster," *Mamm. Genome*, vol. 13, pp. 74–79, Feb. 2002.

[142] G. R. Rompala, V. Zsiros, S. Zhang, S. M. Kolata, and K. Nakazawa, "Contribution of NMDA receptor hypofunction in prefrontal and cortical excitatory neurons to schizophrenia-like phenotypes," *PLoS One*, vol. 8, p. e61278, Apr. 2013.

[143] M. Martin, "Cutadapt removes adapter sequences from high-throughput sequencing reads," *EMBnet.journal*, vol. 17, pp. 10–12, May 2011.

[144] F. Krueger and S. R. Andrews, "Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications," *Bioinformatics*, vol. 27, pp. 1571–1572, June 2011.

[145] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup, "The sequence Alignment/Map format and SAMtools," *Bioinformatics*, vol. 25, pp. 2078–2079, Aug. 2009.

[146] M. A. Urich, J. R. Nery, R. Lister, R. J. Schmitz, and J. R. Ecker, "MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing," *Nat. Protoc.*, vol. 10, pp. 475–483, Mar. 2015.

[147] H. Li, "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data," *Bioinformatics*, vol. 27, pp. 2987–2993, Nov. 2011.

[148] D. Tsafrir, I. Tsafrir, L. Ein-Dor, O. Zuk, D. A. Notterman, and E. Domany, "Sorting points into neighborhoods (SPIN): data analysis and visualization by ordering distance matrices," *Bioinformatics*, vol. 21, pp. 2301–2308, May 2005.

[149] L. Hubert and P. Arabie, "Comparing partitions," *J. Classification*, vol. 2, pp. 193–218, Dec. 1985.

[150] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *J. Mach. Learn. Res.*, vol. 11, no. Oct, pp. 2837–2854, 2010.

[151] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 1, pp. 224–227, Feb. 1979.

[152] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987.

[153] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Commun. Stat. Simul. Comput.*, vol. 3, pp. 1–27, Jan. 1974.

[154] B. Li and C. N. Dewey, "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome," *BMC Bioinformatics*, vol. 12, p. 323, Aug. 2011.

[155] J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J.-Y. Tinevez, D. J. White, V. Hartenstein, K. Eliceiri, P. Tomancak, and A. Cardona, "Fiji: an open-source platform for biological-image analysis," *Nat. Methods*, vol. 9, pp. 676–682, June 2012.

[156] T. Daley and A. D. Smith, "Predicting the molecular complexity of sequencing libraries," *Nat. Methods*, vol. 10, pp. 325–327, Apr. 2013.

[157] C. Luo, M. A. Lancaster, R. Castanon, J. R. Nery, J. A. Knoblich, and J. R. Ecker, "Cerebral organoids recapitulate epigenomic signatures of the human fetal brain," *Cell Rep.*, vol. 17, pp. 3369–3384, Dec. 2016.

[158] E. Wingender, T. Schoeps, and J. Dönitz, "TFClass: an expandable hierarchical classification of human transcription factors," *Nucleic Acids Res.*, vol. 41, pp. D165–70, Jan. 2013.

[159] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu, "Model-based analysis of ChIP-Seq (MACS)," *Genome Biol.*, vol. 9, p. R137, Sept. 2008.